J. Arenas-García, J.Cid-Sueiro, V. Gómez-Verdejo, M. Lázaro-Gredilla, E. Parrado-Hernández

Tratamiento Digital de la Información

– Notas Introductorias –

August 26, 2015

Springer

Contents

1 Estimación analítica 1
1.1 Visión general de los problemas de estimación 1
1.1.1 Estimadores de parámetro determinista y de variable aleatoria 2
1.1.2 Estimación analítica y máquina 3
1.2 Diseño analítico de estimadores de variable aleatoria. Teoría bayesiana de la estimación 4
1.2.1 Modelado estadístico de los problemas de estimación 4
1.2.2 Funciones de coste 5
1.2.3 Coste medio. 7
1.2.4 Estimador bayesiano. 8
1.3 Estimadores bayesianos de uso frecuente 10
1.3.1 Estimador de mínimo error cuadrático medio (MMSE) 10
1.3.2 Estimador de mínimo error absoluto (MAD) 11
1.3.3 Estimador de máximo a posteriori (MAP) 13
1.4 Estimación de máxima verosimilitud 15
1.5 Estimación con distribuciones gaussianas 18
1.5.1 Caso unidimensional 19
1.5.2 Caso con variables multidimensionales 22
1.6 Estimación con restricciones 23
1.6.1 Principios generales 23
1.6.2 Estimación lineal de mínimo error cuadrático medio 24
1.7 Caracterización de estimadores 29
1.7.1 Sesgo y varianza de estimadores de parámetros deterministas 30
1.7.2 Sesgo y varianza de estimadores de variables aleatorias 33
1.8 Apéndices 34
1.8.1 Casos particulares gaussianos 34
1.8.2 Principio de Ortogonalidad. Interpretación geométrica 36
1.9 Problemas 38

2 Aprendizaje Máquina 41
2.1 Principios generales del aprendizaje máquina 41
2.2 Métodos Paramétricos y no Paramétricos 42
2.3 Estimación Máquina No Paramétrica: Método del vecino más próximo 43
2.4 Estimación Máquina Paramétrica: Regresión de Mínimos Cuadrados 43
2.4.1 Modelos Semilineales 43
2.5 Generalización 44

3 Decisión analítica 45
3.1 Introducción al problema de decisión 45
3.1.1 Regiones de decisión 47
3.1.2 Diseño de decisores 47
3.2 Diseño analítico de decisores 48
3.2.1 Modelado estadístico de los problemas de decisión 48
3.2.2 Riesgo 49
3.2.3 Teoría bayesiana de la decisión 51
3.2.4 Decisión ML 54
3.3 Decisores binarios 55
3.3.1 Riesgo de un decisor binario 56
3.3.2 Función discriminante 57
3.3.3 Decisores binarios de mínimo riesgo 58
3.3.4 Decisor ML 59
3.3.5 Decisores no Bayesianos 62
3.4 El caso Gaussiano 67
3.4.1 Varianzas iguales 69
3.4.2 Medias nulas 70
3.5 Apéndices 71
3.5.1 Diseño analítico de decisores con costes dependientes de la observación 71
3.6 Problemas 75

4 Decisión máquina 79
4.1 Diseño de clasificadores bajo enfoque máquina 79
4.1.1 Estimación paramétrica ML para clasificación 80

5 Filtrado Lineal 83
5.1 Introducción 83
5.2 El problema de filtrado 83
5.3 Solución ML 84
5.4 Solución Bayesiana 85
5.4.1 Predicción probabilística de la salida del filtro 86
5.5 Cálculo online 86
5.5.1 Solución Bayesiana 87
5.5.2 Solución ML 87

5.6 Filtro de Wiener 87
5.7 Problemas 88

6 Soluciones de los problemas 89
6.1 Problemas del Capítulo 1 89
6.2 Problemas del Capítulo 3 91
6.3 Problemas del Capítulo 5 94

References 95

1

Estimación analítica

1.1 Visión general de los problemas de estimación

El diseño de un estimador consiste en construir una función real que, a partir del valor de unas determinadas variables de observación, proporcione predicciones acerca de una variable (o vector) objetivo. A modo de ejemplo, considérese la producción de energía en una planta nuclear. Con el fin de maximizar el beneficio de explotación resulta muy deseable adecuar la generación de energía a la demanda real, ya que la capacidad de almacenamiento de la energía no consumida es muy limitada. Obviamente, la demanda energética está muy relacionada con determinados factores, tales como la hora del día, la época del año, la temperatura actual, etc. Por lo tanto, en este contexto está muy generalizado el diseño de modelos de estimación que, a partir de variables de fácil acceso, proporcionan predicciones sobre el consumo energético.

La Figura 1.1 representa el proceso completo de un sistema de estimación. Habitualmente, el resultado de la estimación lleva aparejada una cierta actuación (por ejemplo, generar una determinada cantidad de energía), y los errores en que se incurre al realizar la estimación acarrean determinadas penalizaciones. En este sentido, el objetivo perseguido en el diseño de un estimador suele ser la minimización de dicha penalización (o la maximización de un beneficio) cuando el estimador se utiliza repetidas veces.

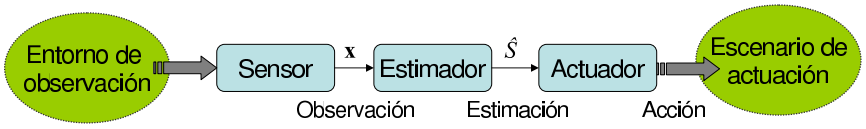


Fig. 1.1. Diagrama de bloques de un sistema de estimación.

Introducimos a continuación la notación que utilizaremos a lo largo del presente capítulo de estimación:

- Denotaremos el valor real de la variable a estimar como s o S, dependiendo de que dicha variable tenga carácter determinista o aleatorio. Si se trata de la estimación de un vector, se denotará como $\mathbf{s}$ o $\mathbf{S}$.
- Incluiremos en un vector aleatorio $\mathbf{X}$ todas aquellas observaciones que representan la información utilizada para cada aplicación concreta del estimador, recogida a través de *sensores* que exploran el escenario lógico (e.g., día del año) y/o físico (e.g., medida de temperatura) en que se lleva a cabo el proceso de estimación. Nótese que el vector de observaciones tiene siempre carácter aleatorio, independientemente del carácter de la variable a estimar. Nótese también que para que la tarea de estimación de s o S a partir de $\mathbf{X}$ tenga sentido, es necesario que exista alguna relación estadística entre ellos.
- El módulo de estimación implementa una función de salida real, $\hat{S} = f(\mathbf{X})$, siendo $f(\cdot)$ la función de estimación. Es habitual referirse a dicha función simplemente como *estimador*, y a su salida como *estimación*. Una característica fundamental del estimador es el carácter determinista de la función $f(\cdot)$, es decir, para un valor dado $\mathbf{x}$ el estimador proporcionará siempre la misma salida. No obstante lo anterior, cuando el argumento de la función es un vector aleatorio, la salida del estimador es una variable aleatoria independientemente de que la variable a estimar sea aleatoria o determinista, y por lo tanto la denotaremos con letra mayúscula.
- El módulo *actuador* llevará a cabo unas u otras actuaciones en función del resultado del proceso de estimación, actuando sobre su entorno. Dado que es de esperar que el estimador incurra en un determinado error en cada aplicación, la actuación será subóptima (i.e., diferente a la que se habría llevado a cabo de conocer de forma exacta el valor de s o S), lo que acarreará un determinado coste (o, alternativamente, un beneficio que conviene maximizar).

Conviene indicar, por último, que en ocasiones el contexto sugerirá cambiar la notación, empleando otros nombres para denotar la variable objetivo y/o las observaciones. Un primer ejemplo de esto se tiene en el Ejemplo 1.1 descrito más abajo.

1.1.1 Estimadores de parámetro determinista y de variable aleatoria

En la exposición previa se ha mencionado que la variable a estimar puede tener carácter determinista o aleatorio. Dicha diferencia no es trivial y tiene importantes consecuencias tanto en el diseño de los correspondientes estimadores, como en la evaluación de sus prestaciones. Por este motivo, una de las primeras reflexiones que han de hacerse a la hora de resolver un problema de estimación es precisamente acerca del carácter aleatorio o determinista de la variable a estimar.

A modo de ejemplo se describen en esta subsección dos casos de estimación, cada uno de ellos correspondiente a un tipo diferente de variable a estimar.

Example 1.1 (Estimación de parámetro determinista). Se desea transmitir información a través de un canal de comunicaciones que introduce ruido blanco y gaussiano, con media nula y varianza desconocida. Para disponer de una caracterización más completa de dicho canal se desea diseñar un estimador de la varianza del ruido basado en la observación de l observaciones independientes del mismo. En este caso, que será resuelto más adelante, la variable a estimar v es determinista, pero puede observarse que el conjunto de observaciones para la estimación $\mathbf{X}$ es un vector aleatorio (son muestras de ruido gaussiano) cuya distribución depende del valor de la varianza. Por tanto, el objetivo será construir una función de la forma

$$\hat{V} = f(\mathbf{X})$$

siendo $\mathbf{X} = [X^{(1)}, X^{(2)}, \ldots, X^{(l)}]$.

Resulta obvio que el conjunto de observaciones permite extraer información acerca del valor real de v. Así, por ejemplo, y dado que la media del ruido es nula, el estimador debería proporcionar valores mayores cuanto mayores fuesen los valores absolutos de las muestras de ruido observadas.

Example 1.2 (Estimación de variable aleatoria). Considérese la tasación de bienes inmuebles. Se desea conocer el precio de mercado S de una vivienda de 3 dormitorios situada en la zona centro de Leganés. Una empresa tasadora conoce a priori que la distribución de los precios de mercado de los inmuebles de ese tipo sigue una determinada distribución probabilística (es decir, se conoce la distribución de precios $p_S(s)$). No obstante, si se desea una estimación más precisa, podría construirse un estimador que tuviese además en cuenta el número de metros cuadrados de la vivienda y de su garaje asociado, distancia de dicha vivienda a la Universidad Carlos III de Madrid, horas de sol que recibe la vivienda, etc. Dichos datos componen un vector de observaciones $\mathbf{X}$ correlacionado con el precio de la vivienda, y por tanto pueden ser utilizados para construir una función de estimación del precio de la forma

$$\hat{S} = f(\mathbf{X})$$

siendo $\mathbf{X} = [\text{m}^2 \text{ de la vivienda}, \text{distancia a UC3M en metros}, \ldots]$.

Nótese que un modelado conjunto de las observaciones y la variable a estimar requeriría conocer $p_{\mathbf{X},S}(\mathbf{x}, s)$, y que tanto esta probabilidad conjunta como la marginal de la variable aleatoria S no pueden ser definidas para el caso en que la variable a estimar tiene carácter determinista. Éste es el rasgo diferenciador fundamental de ambos tipos de problemas de estimación, y la causa de los diferentes planteamientos que para ellos habrán de realizarse en el capítulo.

1.1.2 Estimación analítica y máquina

El diseño de un estimador debe tener en cuenta la relación que existe entre la variable que se desea estimar y las observaciones que se utilizarán como argumento de

entrada del estimador. Según cómo venga dada dicha información, consideraremos dos familias principales de procedimientos de diseño:

- Métodos **analíticos**: se basan en la disponibilidad de cierta información estadística que relaciona observaciones y valor a estimar. El tipo de información requerida para el diseño del estimador varía en función de cuál sea el tipo de estimador que se desea construir (por ejemplo, según sea el criterio de diseño). En general, esta aproximación analítica resulta posible cuando la naturaleza del problema hace posible determinar un modelo probabilístico de las variables involucradas.
- Métodos **máquina**: se basan en la disponibilidad de un conjunto *etiquetado* de datos de entrenamiento, i.e., un conjunto de pares $\{\mathbf{x}^{(k)}, s^{(k)}\}_{k=1}^{l}$. Este conjunto de datos proporciona información acerca de cuál sería la salida deseada del sistema para diferentes valores de las observaciones de entrada. De esta manera, resulta posible partir de una forma paramétrica para la función de estimación $f(\cdot)$, y ajustar los valores de los parámetros de manera que el comportamiento del estimador en los datos de entrenamiento sea el deseado. Nótese, no obstante, que el objetivo del estimador construido es que sea capaz de proporcionar estimaciones acertadas cuando sea aplicado a nuevos datos no vistos durante el entrenamiento. A esta propiedad se la conoce como *capacidad de generaralización* del estimador.

Finalmente, conviene mencionar que existe una tercera vía en la que el conjunto de datos de entrenamiento se utiliza para estimar la información probabilística necesaria para un diseño de tipo analítico. A este tipo de procedimientos se los conoce como métodos **semianalíticos**.

En lo que resta de capítulo se considerarán técnicas para el diseño analítico y máquina de estimadores. En primer lugar, se presentan los conceptos fundamentales para un diseño analítico óptimo, prestando una especial atención a los modelos de estimación lineales en sus parámetros, y al imporante caso en que las variables involucradas tienen carácter gaussiano. Presentaremos, además, criterios que permiten evaluar ciertas propiedades de los estimadores. La parte final del capítulo considera algunas técnicas importantes para el diseño máquina de estimadores, contexto en el que se presentarán algunos conceptos como los modelos semilineales y las técnicas de validación cruzada.

1.2 Diseño analítico de estimadores de variable aleatoria. Teoría bayesiana de la estimación

1.2.1 Modelado estadístico de los problemas de estimación

Antes de abordar el propio diseño de los estimadores, recogemos en esta subsección las distintas funciones de probabilidad que caracterizan estadísticamente la relación existente entre observaciones y variable a estimar:

- En primer lugar, la **verosimilitud** de la variable S viene dada por $p_{\mathbf{X}|S}(\mathbf{x}|s)$, y caracteriza probabilísticamente la generación de las observaciones para cada valor concreto de la variable a estimar.
  En el caso en que la variable a estimar es determinista no tiene sentido condicionar la distribución de probabilidad de las observaciones al valor de s, por lo que lo estrictamente correcto sería denotar la densidad de probabilidad de las observaciones simplemente como $p_{\mathbf{X}}(\mathbf{x})$. No obstante, nótese que para que el problema de estimación tenga sentido, dicha densidad de probabilidad de $\mathbf{X}$ ha de ser diferente según sea el valor real del parámetro determinista. Por este motivo, en ocasiones abusaremos la notación y denotaremos dicha dependencia de las observaciones con s como $p_{\mathbf{X}|s}(\mathbf{x}|s)$, refiriéndonos a dicha densidad de probabilidad como la verosimilitud de s.
- Únicamente en el caso en que la variable a estimar sea aleatoria, podemos definir además densidades de probabilidad sobre S:
  - **Distribución marginal o a priori** de S: $p_S(s)$
  - **Distribución conjunta** de $\mathbf{X}$ y S: $p_{\mathbf{X},S}(\mathbf{x}, s) = p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)$
  - **Distribución a posteriori** de S: $p_{S|\mathbf{X}}(s|\mathbf{x})$.

Es importante resaltar que la información disponible para el diseño del estimador puede ser diferente en cada situación concreta. Una situación habitual, por estar relacionada con el propio proceso físico de generación de las observaciones, es aquélla en la que se dispone de la verosimilitud y de la distribución marginal de S. Nótese que a partir de ellas el cálculo de la distribución conjunta es inmediato, y que dicha distribución conjunta proporciona el modelado estadístico más completo que puede tenerse en un problema de estimación.

Asimismo, es frecuente que el diseño analítico de estimadores requiera el uso de la distribución a posteriori de la variable a estimar, $p_{S|\mathbf{X}}(s|\mathbf{x})$, que indica qué valores de S concentran mayor o menor probabilidad para cada valor concreto del vector de observaciones. Para el cálculo de dicha distribución el Teorema de Bayes resulta ser una herramienta de gran utilidad, ya que permite obtener dicha probabilidad a partir de la distribución a priori de S y de su verosimilitud que, como hemos comentado, suelen ser más accesibles:

$$p_{S|\mathbf{X}}(s|\mathbf{x}) = \frac{p_{\mathbf{X},S}(\mathbf{x}, s)}{p_{\mathbf{X}}(\mathbf{x})} = \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{\int p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)ds} \tag{1.1}$$

Por último, hay que resaltar que según el estimador que se pretenda implementar la información requerida para el diseño puede ser sustancialmente menor que la utilizada para un modelado estadístico completo del problema de estimación. Así, por ejemplo, veremos que para el cálculo de ciertos estimadores resultará suficiente el conocimiento de ciertos momentos estadísticos de la distribución a posteriori de S.

1.2.2 Funciones de coste

El diseño de un estimador requiere algún criterio objetivo. En nuestro caso, consideraremos que dicho criterio puede materializarse en forma de alguna función cuyo

valor perseguimos maximizar o minimizar. Hacemos notar, no obstante, que existen estrategias de diseño que caen fuera de este enfoque.

En el caso concreto de estimación de variable aleatoria, es frecuente definir una función de coste que mide la discrepancia entre el valor real y el estimado de la variable S. Dicho coste está asociado a la penalización que conlleva la aplicación de dicho estimador según el modelo que describimos en la Sección 1.1 de este capítulo. Aceptando que $c(S, \hat{S})$ mide el coste[1], un criterio frecuente de diseño consiste en la minimización de dicho coste en un sentido estadístico, es decir la minimización de la esperanza de la función de coste, lo que equivale a minimizar el coste promedio que se obtendría al realizar un número infinitamente alto de experimentos.

Dado que la función de coste está asociada a una penalización cuyo origen está en la discrepancia entre el valor real y el estimado de S, es frecuente aceptar que $c(s, \hat{s}) \geq 0$, verificándose la igualdad cuando $s = \hat{s}$. Alternativamente, puede definirse una función de beneficio cuyo valor medio ha de ser maximizado. Además, es frecuente que la función de coste no dependa de los valores concretos de s y $\hat{s}$, sino del error de estimación que se define como la diferencia entre ambas, $e = s - \hat{s}$, en cuyo caso tenemos que $c(s, \hat{s}) = c(s - \hat{s}) = c(e)$, y el objetivo de diseño será la minimización de $\mathbb{E}\{c(E)\}$, donde $\mathbb{E}$ denota el operador de esperanza matemática.

A modo de ejemplo, algunas funciones de coste de uso frecuente en el diseño de estimadores son las siguientes:

- Coste cuadrático: $c(e) = e^2$.
- Valor absoluto del error: $c(e) = |e|$.
- Error cuadrático relativo: $c(s, \hat{s}) = \frac{(s-\hat{s})^2}{s^2}$
- Entropía cruzada: $c(s, \hat{s}) = -s \ln \hat{s} - (1 - s) \ln(1 - \hat{s})$, para $s, \hat{s} \in [0, 1]$

Example 1.3 (Error cuadrático medio). Supongamos que X es una observación ruidosa de S, de tal modo que

$$X = S + R \tag{1.2}$$

siendo S una variable aleatoria de media 0 y varianza 1, y R una variable aleatoria gaussiana, independiente de S, de media 0 varianza v. Considerando el estimador $\hat{S} = X$, el coste cuadrático medio es

$$\mathbb{E}\{(S - \hat{S})^2\} = \mathbb{E}\{(S - X)^2\} = \mathbb{E}\{R^2\} = v \tag{1.3}$$

El coste absoluto será

$$\begin{aligned} \mathbb{E}\{|S - \hat{S}|\} = \mathbb{E}\{|R|\} &= \int_{-\infty}^{\infty} |r| \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr \\ &= 2 \int_{0}^{\infty} r \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{r^2}{2v}\right) dr = \sqrt{\frac{2v}{\pi}} \end{aligned} \tag{1.4}$$

[1] Nótese que la función de coste se denota con una c minúscula por ser una función de carácter determinista, i.e., para unos valores fijos de s y $\hat{s}$ el coste siempre toma el mismo valor. Sin embargo, al igual que ocurría con la función de estimación, la aplicación de dicha función sobre variables aleatorias dará lugar a otra variable aleatoria, i.e., $C = c(S, \hat{S})$.

En general, la minimización de cada coste dará lugar a un estimador diferente que será óptimo respecto del coste utilizado en el diseño. Nótese, no obstante la diferencia fundamental que existe entre funciones de coste y estimadores. A pesar de nuestra discusión previa en la que se indica que es habitual diseñar estimadores que son óptimos respecto de una función de coste determinada, resulta completamente viable calcular el coste medio de dicho estimador respecto de cualquier otra función de coste diferente de la empleada para el diseño. A modo de ejemplo, podríamos estar interesados en conocer el coste absoluto medio que resulta de la aplicación del estimador de mínimo error cuadrático medio.

Example 1.4 (Funciones de coste de variables aleatorias multidimensionales). TBD

1.2.3 Coste medio.

De forma general, el coste medio de un estimador viene dado por

$$\mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}}\int_{s} c(s,\hat{s}) p_{S,\mathbf{X}}(s,\mathbf{x}) ds d\mathbf{x} \tag{1.5}$$

donde debe tenerse en cuenta que $\hat{s}$ es, en general, función de $\mathbf{x}$. El coste medio constituye una medida de las prestaciones de un decisor, y por lo tanto proporciona un criterio para comparar dos estimadores cualesquiera.

Example 1.5 (Cálculo del coste medio global). Supongamos que la distribución conjunta de S y X está dada por

$$p_{S,X}(s,x) = \begin{cases} \frac{1}{x}, & 0 < s < x < 1 \\ 0, & \text{resto} \end{cases} \tag{1.6}$$

Consideremos dos estimadores $\hat{S}_1 = \frac{1}{2}X$ y $\hat{S}_2 = X$. ¿Cuál es mejor estimador desde el punto de vista del coste cuadrático? Para averiguarlo, calcularemos el error cuadrático medio para ambos estimadores. Sabiendo que, para cualquier w,

$$\begin{aligned} \mathbb{E}\{(S-wX)^2\} &= \int_0^1 \int_0^x (s-wx)^2 p_{S,X}(s,x) ds dx \\ &= \int_0^1 \int_0^x (s-wx)^2 \frac{1}{x} ds dx \\ &= \int_0^1 \left(\frac{1}{3} - w + w^2\right) x^2 dx \\ &= \frac{1}{3}\left(\frac{1}{3} - w + w^2\right) \end{aligned} \tag{1.7}$$

Tomando $w = 1/2$ resulta

$$\mathbb{E}\{(S-\hat{S}_1)^2\} = \mathbb{E}\{(S-\frac{1}{2}X)^2\} = \frac{1}{3}\left(\frac{1}{3}-\frac{1}{2}+\frac{1}{4}\right) = \frac{1}{36} \tag{1.8}$$

Alternativamente, tomando $w = 1$ se obtiene

$$\mathbb{E}\{(S-\hat{S}_2)^2\} = \mathbb{E}\{(S-X)^2\} = \frac{1}{3}\left(\frac{1}{3}-1+1\right) = \frac{1}{9} \tag{1.9}$$

Por tanto, desde el punto de vista del error cuadrático medio, $\hat{S}_1$ es mejor estimador que $\hat{S}_2$

1.2.4 Estimador bayesiano.

Cabe preguntarse, para un coste y una distribución dadas, cuál es el mejor estimador posible. Podemos averiguarlo teniendo en cuenta que, de forma general, el coste medio en (1.5) puede expresarse como

$$\begin{aligned}\mathbb{E}\{c(S,\hat{S})\} &= \int_{\mathbf{x}}\int_{s} c(s,\hat{s})p_{S|\mathbf{X}}(s|\mathbf{x})ds\; p_{\mathbf{X}}(\mathbf{x})\mathbf{dx} = \\ &= \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\}p_{\mathbf{X}}(\mathbf{x})\mathbf{dx}.\end{aligned} \tag{1.10}$$

La última línea de esta ecuación muestra que una estrategia que permite minimizar el error de estimación global consiste en la minimización del error medio para cada posible valor del vector de observaciones, $\mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\}$, al que nos referiremos como coste medio a posteriori o coste medio dado $\mathbf{X}$. Por tanto, ambas estrategias (minimización de la esperanza del error para todo S y $\mathbf{X}$, o condicionado al valor de $\mathbf{X}$) son en principio equivalentes de cara a obtener el estimador óptimo asociado a una función de coste determinada.

Se define el Estimador bayesiano asociado a una función de coste como aquél que minimiza (1.10), es decir:

$$\hat{s}^* = \underset{\hat{s}}{\operatorname{argmin}}\;\; \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\} \tag{1.11}$$

donde $\hat{s}^*$ es el Estimador bayesiano. De acuerdo a nuestra discusión previa, el Estimador bayesiano minimiza también el coste esperado en un sentido global, i.e., para todo S y $\mathbf{X}$. Nótese, sin embargo, que para su diseño resulta más útil la expresión (1.11) que la minimización directa del coste global

$$\mathbb{E}\{c(S,\hat{S})\} = \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\}p_{\mathbf{X}}(\mathbf{x})\mathbf{dx} \tag{1.12}$$

ya que el cálculo de la integral en $\mathbf{x}$ requeriría conocer de antemano la relación que existe entre $\hat{s}$ y $\mathbf{x}$, lo que constituye precisamente el objetivo del problema de diseño del estimador.

Example 1.6 (Cálculo de un estimador de mínimo coste cuadrático medio). Continuado el ejemplo 1.5, podemos calcular la distribución a posteriori de S mediante

$$p_{S|X}(s|x) = \frac{p_{S,X}(s,x)}{p_X(x)}. \tag{1.13}$$

Sabiendo que

$$p_X(x) = \int_0^1 p_{S,X}(s,x)ds = \int_0^x \frac{1}{x}ds = 1, \tag{1.14}$$

resulta

$$p_{S|X}(s|x) = \begin{cases} \frac{1}{x}, & 0 < s < x < 1 \\ 0, & \text{resto} \end{cases} \tag{1.15}$$

El coste medio dada la observación vendrá dado por

$$\begin{aligned} \mathbb{E}\{c(S,\hat{s})|X=x\} &= \mathbb{E}\{(S-\hat{s})^2|X=x\} \\ &= \int_0^1 (s-\hat{s})^2 p_{S|X}(s|x)ds \\ &= \frac{1}{x}\int_0^x (s-\hat{s})^2 ds \\ &= \frac{1}{x}\left(\frac{(x-\hat{s})^3}{3} + \frac{\hat{s}^3}{3}\right) \\ &= \frac{1}{3}x^2 - \hat{s}x + \hat{s}^2. \end{aligned} \tag{1.16}$$

Como función de $\hat{s}$, el coste medio condicionado a la observación es un polinomio de segundo grado, cuyo mínimo puede calcularse de modo inmediato por derivación. Siendo

$$\frac{d}{d\hat{s}}\mathbb{E}\{c(S,\hat{s})|X=x\} = -x + 2\hat{s}, \tag{1.17}$$

el estimador de mínimo coste cuadrático medio será

$$\hat{s}^* = \frac{1}{2}x, \tag{1.18}$$

que coincide con el estimador $\hat{S}_1$ del ejemplo 1.5. Por tanto, $\hat{S}_1$ es el mejor estimador posible desde el punto de vista del coste cuadrático medio.

De acuerdo con (1.11) podemos concluir que, con independencia del coste que se pretenda minimizar, el conocimiento de la distribución a posteriori de S dado $\mathbf{X}$, $p_{S|\mathbf{X}}(s|\mathbf{x})$, resulta suficiente para el diseño del Estimador bayesiano óptimo. Como ya se ha comentado, dicha distribución es frecuentemente calculada a partir de la verosimilitud de S y de su distribución a priori utilizando el Teorema de Bayes, lo que de hecho constituye el origen de la denominación de estos estimadores.

1.3 Estimadores bayesianos de uso frecuente

En esta sección se presentan algunos de los estimadores bayesianos de uso más común. Para su cálculo, procederemos a la minimización del coste medio dado $\mathbf{X}$ (coste medio a posteriori) para distintas funciones de coste.

1.3.1 Estimador de mínimo error cuadrático medio (MMSE)

El estimador de mínimo error cuadrático medio (*Minimum Mean Square Error*, MMSE) es el asociado a la función de coste $c(e) = e^2 = (s - \hat{s})^2$, y por lo tanto queda caracterizado por

$$\hat{s}_{\text{MMSE}} = \underset{\hat{s}}{\operatorname{argmin}} \ \mathbb{E}\{c(S, \hat{s}) | \mathbf{X} = \mathbf{x}\} = \tag{1.19}$$

$$= \underset{\hat{s}}{\operatorname{argmin}} \int_s (s - \hat{s})^2 p_{S|\mathbf{X}}(s|\mathbf{x}) ds \tag{1.20}$$

La Figura 1.2 ilustra el problema de diseño del estimador de mínimo error cuadrático medio. El coste medio a posteriori se puede obtener integrando en s la función que resulta del producto de la función de coste y de la densidad de probabilidad a posteriori de S. El argumento para la minimización es $\hat{s}$, lo que permite desplazar la gráfica correspondiente a la función de coste (representada con trazo discontinuo) de forma que el resultado de dicha integral sea mínimo.

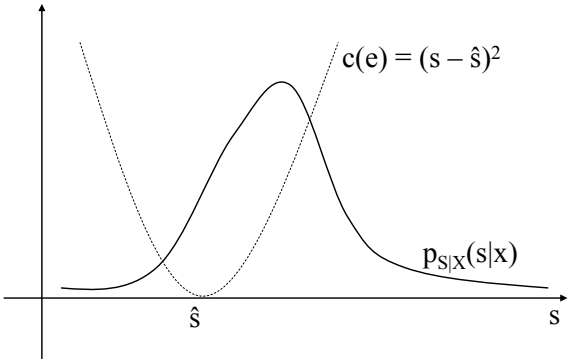


Fig. 1.2. Representación gráfica del proceso de cálculo del coste cuadrático medio a posteriori para un valor genérico $\hat{s}$.

El valor de $\hat{s}_{\text{MMSE}}$ puede obtenerse de forma analítica tomando la derivada del coste medio a posteriori e igualando el resultado a 0. El cálculo de la derivada no

plantea ninguna dificultad ya que la derivada y la integral pueden conmutarse (se integra respecto de s y se deriva respecto de $\hat{s}$):

$$\left.\frac{d\mathbb{E}\{(S-\hat{s})^2|\mathbf{X}=\mathbf{x}\}}{d\hat{s}}\right|_{\hat{s}=\hat{s}_{\text{MMSE}}} = -2\int_s (s-\hat{s}_{\text{MMSE}})p_{S|\mathbf{X}}(s|\mathbf{x})ds = 0 \qquad (1.21)$$

Teniendo en cuenta que la integral que aparece en (1.21) debe anularse, y utilizando el hecho de que $\int p_{S|\mathbf{X}}(s|\mathbf{x})ds = 1$, resulta sencillo demostrar que el estimador de mínimo error cuadrático medio de S viene dado por

$$\hat{s}_{\text{MMSE}} = \int s\; p_{S|\mathbf{X}}(s|x)ds = \mathbb{E}\{S|\mathbf{X}=\mathbf{x}\} \qquad (1.22)$$

En otras palabras, el estimador de mínimo error cuadrático medio de S es la media a posteriori de S dado $\mathbf{X}$, i.e., la media de $p_{S|\mathbf{X}}(s|\mathbf{x})$.

Exercise 1.7. Compruebe que la expresión (1.22) efectivamente constituye un mínimo del coste medio dado $\mathbf{X}$, mediante el cálculo de la derivada segunda de $\mathbb{E}\{c(S,\hat{s})|\mathbf{X}=\mathbf{x}\}$.

Example 1.8 (Cálculo directo del estimador MMSE). De acuerdo con (1.22), el estimador de mínimo coste cuadrático medio obtenido en 1.5 puede obtenerse alternativamente como

$$\hat{s}_{\text{MMSE}} = \int_0^1 s p_{S|X}(s|x)ds = \int_0^x \frac{s}{x}ds = \frac{1}{2}x \qquad (1.23)$$

que coincide con (1.18).

1.3.2 Estimador de mínimo error absoluto (MAD)

De forma similar a como hemos procedido para el caso del estimador $\hat{s}_{\text{MMSE}}$, podemos calcular el estimador asociado al valor absoluto del error de estimación, $c(e) = |e| = |s-\hat{s}|$. Dicho estimador, al que nos referiremos como estimador de mínimo error absoluto (*Mean Absolute Deviation*, MAD), está caracterizado por

$$\begin{aligned}\hat{s}_{\text{MAD}} &= \underset{\hat{s}}{\text{argmin}}\;\; \mathbb{E}\{|S-\hat{s}|\;|\mathbf{X}=\mathbf{x}\} = \\ &= \underset{\hat{s}}{\text{argmin}} \int_s |s-\hat{s}|\; p_{S|\mathbf{X}}(s|\mathbf{x})ds\end{aligned} \qquad (1.24)$$

Nuevamente, resulta sencillo ilustrar el proceso de cálculo del coste medio a posteriori superponiendo en unos mismos ejes el coste expresado como función de s y

la distribución a posteriori de la variable a estimar (véase la Fig. 1.3). Dicha representación sugiere también la conveniencia de partir la integral en dos tramos correspondientes a las dos ramas de la función de coste:

$$\begin{aligned}\mathbb{E}\{|S-\hat{s}|\ |\mathbf{X}=\mathbf{x}\} &= \int_{-\infty}^{\hat{s}} (\hat{s}-s)\ p_{S|\mathbf{X}}(s|\mathbf{x})ds + \int_{\hat{s}}^{\infty} (s-\hat{s})\ p_{S|\mathbf{X}}(s|\mathbf{x})ds \\ &= \hat{s}\left[\int_{-\infty}^{\hat{s}} p_{S|\mathbf{X}}(s|\mathbf{x})ds - \int_{\hat{s}}^{\infty} p_{S|\mathbf{X}}(s|\mathbf{x})ds\right] + \\ &\quad + \int_{\hat{s}}^{\infty} s\ p_{S|\mathbf{X}}(s|\mathbf{x})ds - \int_{-\infty}^{\hat{s}} s\ p_{S|\mathbf{X}}(s|\mathbf{x})ds\end{aligned} \tag{1.25}$$

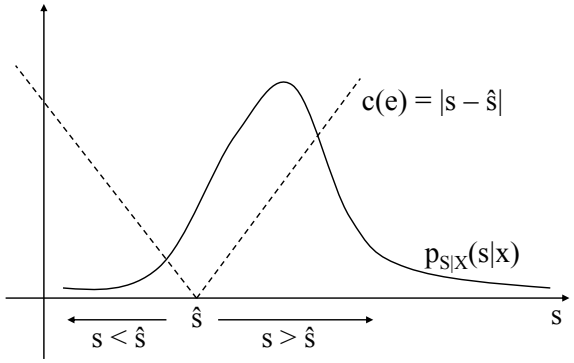


Fig. 1.3. Representación gráfica del proceso de cálculo del coste medio absoluto a posteriori para un valor genérico $\hat{s}$.

El Teorema Fundamental del Cálculo[2] permite obtener la derivada del coste medio a posteriori como

$$\frac{d\mathbb{E}\{|S-\hat{s}|\ |\mathbf{X}=\mathbf{x}\}}{d\hat{s}} = 2F_{S|\mathbf{X}}(\hat{s}|\mathbf{x}) - 1 \tag{1.26}$$

donde $F_{S|\mathbf{X}}(s|\mathbf{x})$ es la función de distribución a posteriori de S dado $\mathbf{X}$. Dado que $\hat{s}_{\text{MAD}}$ representa el mínimo del coste medio, la derivada anterior debe anularse para el estimador, por lo que se ha de verificar que $F_{S|\mathbf{X}}(\hat{s}_{\text{MAD}}|\mathbf{x}) = 1/2$. Dicho de otra manera, el estimador de mínimo error absoluto viene dado por la mediana de $p_{S|\mathbf{X}}(s|\mathbf{x})$:

[2] $\frac{d}{dx}\int_{t_0}^{x} g(t)dt = g(x)$.

$$\hat{s}_{\text{MAD}} = \text{mediana}\{S|\mathbf{X} = \mathbf{x}\} \tag{1.27}$$

Recuérdese que la mediana de una distribución es el punto que separa dicha distribución en dos regiones que acaparan la misma probabilidad, por lo que el estimador de mínimo error absoluto medio verificará que

$$P\{S > \hat{s}_{\text{MAD}}\} = P\{S < \hat{s}_{\text{MAD}}\}$$

Example 1.9 (Diseño de estimador de Mínimo Error Absoluto). En el escenario del ejemplo 1.5, la distribucion a posteriori de S dado X es uniforme entre 0 y x, cuya mediana es $x/2$. Por tanto,

$$\hat{s}_{\text{MAD}} = \frac{1}{2}x \tag{1.28}$$

Observe que, en este caso, el estimador MAD coincide con el MMSE obtenido en (1.18). Esto es una consecuencia de la simetría de la distribución a posteriori. En general, ambos estimadores no tienen por qué coincidir.

1.3.3 Estimador de máximo a posteriori (MAP)

Como su propio nombre indica, el estimador de máximo a posteriori (*Maximum a Posteriori*, MAP) se define como el valor de S que maximiza la distribución de probabilidad a posteriori de dicha variable, i.e., el valor de S que concentra mayor densidad de probabilidad para cada valor de la variable observable:

$$\hat{s}_{\text{MAP}} = \underset{s}{\text{argmax}} \; p_{S|\mathbf{X}}(s|\mathbf{x}) \tag{1.29}$$

En sentido estricto, el estimador MAP no es bayesiano, porque no minimiza ningún coste medio. No obstante, si consideramos la función de coste (véase también la Figura 1.4)

$$c_{\Delta}(s - \hat{s}) = \begin{cases} 1 & ; \text{para } |s - \hat{s}| > \Delta \\ 0 & ; \text{para } |s - \hat{s}| < \Delta \end{cases} \tag{1.30}$$

y denotamos por $\hat{s}_{\Delta}$ el estimador bayesiano asociado a la misma, puede comprobarse que $\hat{s}_{\text{MAP}} = \lim_{\Delta \to 0} \hat{s}_{\Delta}$. El estimador MAP es, por tanto, un caso límite de una familia de estimadores bayesianos.

Exercise 1.10. Demuestre que el estimador MAP puede obtenerse como $\hat{s}_{\text{MAP}} = \lim_{\Delta \to 0} \hat{s}_{\Delta}$.

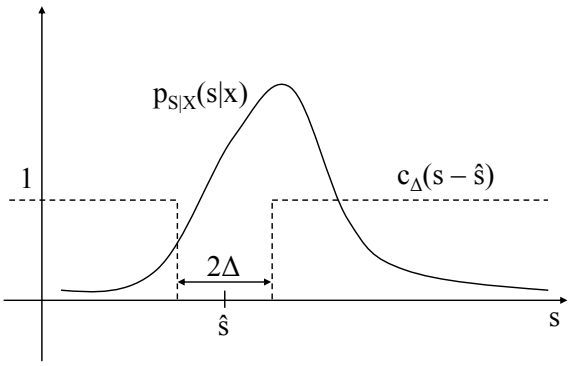


Fig. 1.4. Representación gráfica del proceso de cálculo del coste medio a posteriori para $c_{\Delta}(s-\hat{s})$.

Por otro lado, por motivos prácticos, para la maximización de (1.29) puede ser útil introducir una función auxiliar que simplifique la forma analítica de la función a maximizar. Así, por ejemplo, la definición (1.29) es completamente equivalente a

$$\hat{s}_{\text{MAP}} = \underset{s}{\text{argmax}} \;\; \ln\left[p_{S|\mathbf{X}}(s|\mathbf{x})\right] \tag{1.31}$$

dado que la función logaritmo está definida para todo valor positivo de su argumento y es estrictamente creciente (lo que implica que si $p_{S|\mathbf{X}}(s_1|\mathbf{x}) > p_{S|\mathbf{X}}(s_2|\mathbf{x})$, entonces también $\ln p_{S|\mathbf{X}}(s_1|\mathbf{x}) > \ln p_{S|\mathbf{X}}(s_2|\mathbf{x})$). La introducción de la función logaritmo resultará útil cuando la distribución a posteriori de S dado $\mathbf{X}$ presente productos o exponenciales, ya que transformará productos en sumas y cancelará las exponenciales. De esta manera, el proceso de maximización puede simplificarse considerablemente.

Example 1.11 (Cálculo de un estimador MAP). Suponiendo que

$$p(s|x) = \frac{1}{x^2} s \exp\left(-\frac{s}{x}\right), \qquad x \geq 0, s \geq 0 \tag{1.32}$$

el estimador MAP puede obtenerse maximizando

$$\ln(p(s|x)) = -2\ln(x) + \ln(s) - \frac{s}{x}, \qquad x \geq 0, s \geq 0 \tag{1.33}$$

Dado que $\ln(p(s|x))$ tiende a $-\infty$ en torno a $s = 0$ y $s = \infty$, su máximo debe estar en algún punto intermedio de derivada nula. Derivando respecto a s, resulta

$$\frac{\partial}{\partial s} \ln p(s|x)\bigg|_{s=\hat{s}_{\text{MAP}}} = \frac{1}{\hat{s}_{\text{MAP}}} - \frac{1}{x} = 0, \qquad x \geq 0, s \geq 0 \tag{1.34}$$

Por tanto

$$\hat{s}_{\text{MAP}} = x \tag{1.35}$$

Cuando la distribución a posteriori tiene varios máximos globales, el estimador MAP no es único.

Example 1.12 (Multiplicidad del estimador MAP). En el ejemplo 1.5, la distribucion a posteriori de S dado X es uniforme entre 0 y x. Por tanto, cualquier valor de $s \in [0, x]$ es un estimador MAP.

1.4 Estimación de máxima verosimilitud

Definimos el estimador de máxima verosimilitud (*Maximum Likelihood*, ML) de una variable aleatoria como

$$\hat{s}_{\text{ML}} = \underset{s}{\operatorname{argmax}}\, p_{\mathbf{X}|S}(\mathbf{x}|s) = \underset{s}{\operatorname{argmax}}\, \ln(p_{\mathbf{X}|S}(\mathbf{x}|s)) \tag{1.36}$$

donde se ha indicado que el uso de la función logaritmo (o de alguna otra de propiedades similares) es opcional y no afecta en ningún caso al valor que resulta de la maximización. Es importante resaltar que la maximización de $p_{\mathbf{X}|S}(\mathbf{x}|s)$ ha de realizarse con respecto del valor de s, que no es la variable respecto de la que está definida dicha función de probabilidad.

El estimador de Máxima Verosimilitud no está asociado a la minimización de ningún coste medio a posteriori, y por lo tanto no se considera un estimador bayesiano. De hecho, su aplicación sufre el inconveniente de no tomar en consideración la distribución a priori de la variable aleatoria S. Precisamente, el uso del estimador ML está más justificado en aquellos casos en los dicha información no se encuentra disponible.

El estimador ML coincide con el MAP cuando S presenta distribución uniforme en un rango de valores y, por lo tanto, la aplicación del estimador ML en ausencia de información acerca de la distribución a priori de S equivale a asumir uniformidad para la misma y aplicar el estimador MAP. Para comprobar la equivalencia entre $\hat{s}_{\text{ML}}$ y $\hat{s}_{\text{MAP}}$ cuando $p_S(s)$ es uniforme, no hay más que considerar la relación existente entre la verosimilitud y la distribución a posteriori de S, que según el Teorema de Bayes es

$$p_{S|\mathbf{X}}(s|\mathbf{x}) = \frac{p_{\mathbf{X}|S}(\mathbf{x}|s)p_S(s)}{p_{\mathbf{X}}(\mathbf{x})}$$

Dado que $p_{\mathbf{X}}(\mathbf{x})$ no depende de s y estamos asumiendo que $p_S(s)$ es constante, el valor de s que maximiza el término izquierdo de la igualdad ha de coincidir con el que maximiza la verosimilitud.

Por último, hay que resaltar que, al contrario de lo que ocurría en el caso de estimación bayesiana, la estimación de máxima verosimilitud no precisa de la definición de densidades de probabilidad sobre la variable a estimar y, por lo tanto, puede

ser aplicada tanto en el caso de estimación de variable aleatoria como de parámetro determinista.

Example 1.13 (Estimación ML de Variable Aleatoria). Se desea estimar el valor de una variable aleatoria S a partir de una observación X estadísticamente relacionada con ella. Para el diseño del estimador se conoce únicamente la verosimilitud de S que está dada por

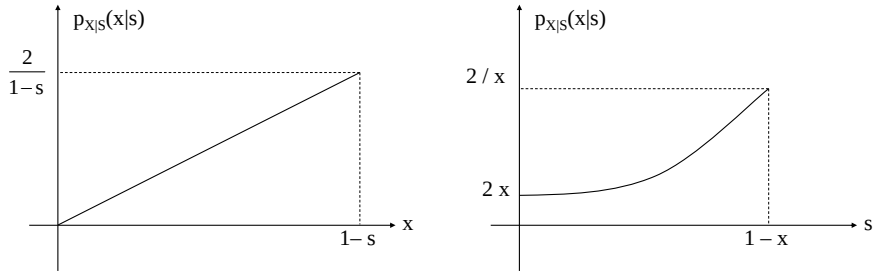
$$p_{X|S}(x|s) = \frac{2x}{(1-s)^2}, \;\; 0 < x < 1-s, \;\; 0 < s < 1 \tag{1.37}$$

Dada la información estadística disponible, se decide construir el estimador ML de S. Para ello, se debe maximizar la verosimilitud anterior con respecto de s. Dicha verosimilitud es una función de densidad de probabilidad de X, tal y como se representa en la Figura 1.5(a), donde se comprueba que la integral de dicha función con respecto de x es unitaria. Sin embargo, para llevar a cabo la maximización que permite encontrar $\hat{s}_{ML}$ resulta de mayor utilidad representar dicha verosimilitud como función de s (Fig. 1.5(b))[3]. A partir de dicha representación gráfica resulta evidente que el estimador buscado es

$$\hat{s}_{\text{ML}} = 1 - x$$

o, alternativamente, si consideramos la aplicación de la función de estimación sobre la variable aleatoria X en lugar de sobre un valor concreto de la misma,

$$\hat{S}_{\text{ML}} = 1 - X$$



(a) (b)

Fig. 1.5. Representación de la función de verosimilitud del Ejericio 1.13 como función de x y de s.

[3] Nótese que la integral respecto de s de $p_{X|S}(x|s)$ no será en general la unidad, ya que dicha función no constituye una densidad de probabilidad de S.

Example 1.14 (Estimación ML de los parámetros de una variable aleatoria gaussiana unidimensional). Se sabe que el peso de los individuos de una familia de moluscos sigue una distribución de tipo gaussiano, cuya media y varianza se desea estimar. Se dispone para la estimación de los pesos de l individuos tomados de forma independiente, $\{X^{(k)}\}_{k=1}^{l}$.

En este caso, nuestro objetivo consiste en construir estimadores de parámetros deterministas, ya que no existe y carece de sentido definir la distribución de probabilidad de la media y la varianza de la distribución gaussiana. La verosimilitud de la media y la varianza, en este caso, consiste simplemente en la distribución de probabilidad de las observaciones, que según establece el enunciado viene dada por:

$$p_X(x) = p_{X|m,v}(x|m,v) = \frac{1}{\sqrt{2\pi v}} \exp\left[-\frac{(x-m)^2}{2v}\right] \tag{1.38}$$

para cada una de las observaciones. Dado que debemos construir el estimador basado en la observación conjunta de l observaciones, necesitaremos calcular la distribución conjunta de todas ellas que, al tratarse de observaciones independientes, se obtiene como producto de las individuales:

$$\begin{aligned} p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v) &= \prod_{k=1}^{l} p_{X|m,v}(x^{(k)}|m,v) \\ &= \frac{1}{(2\pi v)^{l/2}} \prod_{k=1}^{l} \exp\left[-\frac{(x^{(k)}-m)^2}{2v}\right] \end{aligned} \tag{1.39}$$

Los estimadores de máxima verosimilitud de m y de v serán los valores de dichos parámetros que hacen máxima la expresión anterior. La forma analítica de (1.39) sugiere el uso de la función logaritmo para simplificar el proceso de maximización:

$$L = \ln\left[p_{\{X^{(k)}\}|m,v}(\{x^{(k)}\}|m,v)\right] = -\frac{l}{2}\ln(2\pi v) - \frac{1}{2v}\sum_{k=1}^{l}(x^{(k)}-m)^2 \tag{1.40}$$

Para obtener los estimadores de máxima verosimilitud procederemos a derivar (1.40) con respecto de m y de v, y a igualar el resultado con respecto de 0. De esta manera, el sistema de ecuaciones a resolver queda

$$\begin{aligned} \left.\frac{d\,L}{d\,m}\right|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} &= -\frac{1}{v}\sum_{k=1}^{l}(x^{(k)}-m)\Bigg|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} = 0 \\ \left.\frac{d\,L}{d\,v}\right|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} &= -\frac{l}{2v} + \frac{1}{2v^2}\sum_{k=1}^{l}(x^{(k)}-m)^2\Bigg|_{\substack{m=\hat{m}_{\mathrm{ML}} \\ v=\hat{v}_{\mathrm{ML}}}} = 0 \end{aligned} \tag{1.41}$$

La primera de estas ecuaciones permite obtener el estimador de la media de forma sencilla como el promedio muestral de las observaciones, i.e.,

$$\hat{m}_{\text{ML}} = \frac{1}{l}\sum_{k=1}^{l} x^{(k)} \tag{1.42}$$

Por otro lado, podemos despejar el estimador ML de la varianza de la segunda ecuación del sistema, obteniendo

$$\hat{v}_{\text{ML}} = \frac{1}{l}\sum_{k=1}^{l}(x^{(k)} - \hat{m}_{\text{ML}})^2 \tag{1.43}$$

Nótese que, si en lugar de aplicar la función de estimación (de m o de v) sobre unas observaciones concretas lo hiciésemos sobre valores genéricos $\{X^{(k)}\}$, los estimadores podrían ser tratados como variables aleatorias, i.e.,

$$\hat{M}_{\text{ML}} = \frac{1}{l}\sum_{k=1}^{l} X^{(k)} \tag{1.44}$$

$$\hat{V}_{\text{ML}} = \frac{1}{l}\sum_{k=1}^{l}[X^{(k)} - \hat{M}_{\text{ML}}]^2 \tag{1.45}$$

Esto es así porque, a pesar de ser m y v parámetros deterministas, sus estimadores son funciones de las observaciones, y éstas siempre tienen carácter aleatorio.

Exercise 1.15 (Estimación ML de la media de una variable aleatoria gaussiana multidimensional). Demuestre que el estimador ML de la media de una variable gaussiana multidimensional, a partir de l observaciones independientes de la misma, $\{\mathbf{X}^{(k)}\}_{k=1}^{l}$, está dado por el promedio muestral:

$$\hat{\mathbf{m}}_{\text{ML}} = \frac{1}{l}\sum_{k=1}^{l} \mathbf{x}^{(k)}$$

1.5 Estimación con distribuciones gaussianas

En esta sección analizaremos el caso de estimación de variable aleatoria cuando la distribución conjunta de todas las variables implicadas (variable a estimar y variables de observación) es una gaussiana multidimensional. Este caso resulta de especial interés dada la frecuencia con la que dichas distribuciones suelen aparecer en problemas del ámbito de las telecomunicaciones y en otros escenarios. En este caso, puede demostrarse que todas las distribuciones marginales y todas las condicionales son también gaussianas. En concreto, dado que $p_{S|\mathbf{X}}(s|\mathbf{x})$ es gaussiana, puede entenderse que la moda, la media y la mediana de la distribución coinciden, por lo que se verificará $\hat{s}_{\text{MMSE}} = \hat{s}_{\text{MAD}} = \hat{s}_{\text{MAP}}$. Por lo tanto, durante esta sección centraremos nuestra discusión en el cálculo del estimador de mínimo error cuadrático medio.

1.5.1 Caso unidimensional

Consideraremos como punto de partida un caso con variables aleatorias unidimensionales con medias nulas, en el que la distribución conjunta de X y S tiene la siguiente forma:

$$p_{S,X}(s,x) \sim G\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}v_S & \rho\\ \rho & v_X\end{bmatrix}\right) \tag{1.46}$$

siendo ρ la covarianza entre ambas variables aleatorias.

A partir de dicha distribución conjunta podemos obtener cualquier otra distribución que involucre a las variables s y x; en concreto, la distribución a posteriori de S se puede obtener como:

$$\begin{aligned} p_{S|X}(s|x) &= \frac{p_{S,X}(s,x)}{p_X(x)} \\ &= \frac{\frac{1}{2\pi\sqrt{v_X v_S - \rho^2}} \exp\left[-\frac{1}{2(v_X v_S - \rho^2)} \begin{bmatrix}s\\x\end{bmatrix}^T \begin{bmatrix}v_X & -\rho\\ -\rho & v_S\end{bmatrix} \begin{bmatrix}s\\x\end{bmatrix}\right]}{\frac{1}{\sqrt{2\pi v_X}} \exp\left[-\frac{x^2}{2v_X}\right]} \end{aligned} \tag{1.47}$$

donde ha sido necesario calcular la inversa de la matriz de covarianzas de S y X, lo que resulta sencillo al ser dicha matriz de dimensiones 2×2.

Nuestro objetivo para obtener $\hat{s}_{\text{MMSE}}$ consiste en calcular la media de dicha distribución. Sin embargo, un cálculo directo mediante la integración de su producto con s resulta bastante complicado. Sin embargo, dado el carácter conjuntamente gaussiano de S y X, sabemos que la distribución a posteriori de S ha de ser necesariamente gaussiana, definida por sus parámetros (desconocidos) de media y varianza $m_{S|X}$ y $v_{S|X}$, respectivamente, lo que permite reescribir la expresión anterior como:

$$\begin{aligned} &\frac{1}{\sqrt{2\pi v_{S|X}}} \exp\left[-\frac{(s - m_{S|X})^2}{2v_{S|X}}\right] = \\ &\qquad \frac{\frac{1}{2\pi\sqrt{v_X v_S - \rho^2}} \exp\left[-\frac{1}{2(v_X v_S - \rho^2)} \begin{bmatrix}s\\x\end{bmatrix}^T \begin{bmatrix}v_X & -\rho\\ -\rho & v_S\end{bmatrix} \begin{bmatrix}s\\x\end{bmatrix}\right]}{\frac{1}{\sqrt{2\pi v_X}} \exp\left[-\frac{x^2}{2v_X}\right]} \end{aligned} \tag{1.48}$$

Resulta posible descomponer esta igualdad en otras dos asociadas a los factores externos a las exponenciales y a sus argumentos:

$$\frac{1}{\sqrt{2\pi v_{S|X}}} = \frac{\sqrt{2\pi v_X}}{2\pi\sqrt{v_X v_S - \rho^2}} \tag{1.49}$$

$$\frac{(s - m_{S|X})^2}{v_{S|X}} = \frac{1}{v_X v_S - \rho^2} \begin{bmatrix}s\\x\end{bmatrix}^T \begin{bmatrix}v_X & -\rho\\ -\rho & v_S\end{bmatrix} \begin{bmatrix}s\\x\end{bmatrix} - \frac{x^2}{v_X} \tag{1.50}$$

Operando los términos matriciales, la segunda de estas igualdades puede ser reescrita de forma más sencilla como

$$\frac{(s-m_{S|X})^2}{v_{S|X}} = \frac{v_X s^2 + v_S x^2 - 2\rho x s}{v_X v_S - \rho^2} - \frac{x^2}{v_X} \tag{1.51}$$

Nótese que (1.51) supone una igualdad entre dos polinomios en s (y en x). Por lo tanto, los coeficientes de los términos independientes, lineales y cuadráticos en s (i.e., que no dependen de s, o que multiplican a s y s^2) que aparecen en ambos lados de la igualdad deben coincidir. Por lo tanto, y teniendo en cuenta que $m_{S|X}$ no depende de s, se han de verificar las tres igualdades siguientes:

$$\frac{m_{S|X}^2}{v_{S|X}} = \frac{v_S x^2}{v_X v_S - \rho^2} - \frac{x^2}{v_X} \tag{1.52}$$

$$\frac{s\, m_{S|X}}{v_{S|X}} = \frac{\rho x s}{v_X v_S - \rho^2} \tag{1.53}$$

$$\frac{s^2}{v_{S|X}} = \frac{v_X s^2}{v_X v_S - \rho^2} \tag{1.54}$$

Para el cálculo de la media a posteriori, resulta cómodo despejar dicho valor de (1.53) como

$$m_{S|X} = \frac{v_{S|X}\rho x}{v_X v_S - \rho^2} \tag{1.55}$$

Finalmente, el valor de la varianza a posteriori puede extraerse fácilmente de (1.49) o (1.54) como

$$v_{S|X} = \frac{v_X v_S - \rho^2}{v_X} \tag{1.56}$$

Introduciendo este valor en (1.55) se obtiene la expresión que determina el estimador de mínimo error cuadrático medio.

$$\hat{s}_{\text{MMSE}} = m_{S|X} = \frac{\rho}{v_X} x \tag{1.57}$$

Como puede comprobarse, el estimador obtenido tiene carácter lineal.

Exercise 1.16. Generalice el resultado anterior para el caso en que las variables S y X tienen medias no nulas m_S y m_X, respectivamente. Demuestre que en dicho caso, el estimador buscado es

$$\hat{s}_{\text{MMSE}} = m_S + \frac{\rho}{v_X}(x - m_X) \tag{1.58}$$

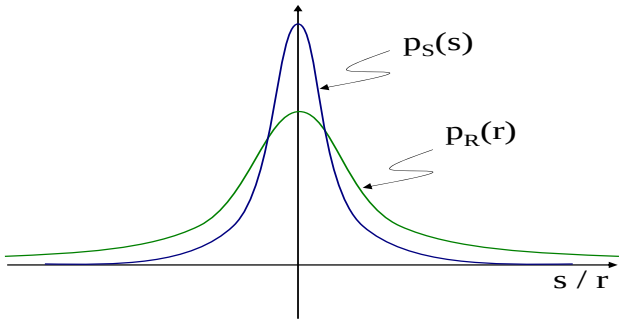


Fig. 1.6. Estimación de variable aleatoria gaussiana S contaminada por ruido gaussiano R.

Example 1.17 (Estimación de señal gaussiana contaminada por ruido gaussiano).
En este ejemplo consideraremos el caso en que la observación se obtiene como suma de la señal a estimar y una componente de ruido independiente de la señal: $X = S + R$. Tanto la señal como el ruido presentan distribuciones gaussianas de medias nulas y varianzas v_S y v_R, respectivamente. La Figura (1.6) representa la situación descrita para un caso con $v_S < v_R$.

De acuerdo con (1.57), para la resolución del problema debemos encontrar la varianza de X y la covarianza entre S y X (ρ). La varianza v_X se obtiene simplemente como la suma de v_S y v_R por ser ambas variables independientes. Para el cálculo de la covarianza podemos proceder como sigue:

$$\rho = \mathbb{E}\{(X-m_X)(S-m_S)\} = \mathbb{E}\{X\ S\} = \mathbb{E}\{(S+R)S\} = \mathbb{E}\{S^2\}+\mathbb{E}\{S\ R\} = v_S \tag{1.59}$$

donde se ha utilizado la independencia de S y R, y el hecho de que todas las variables (incluida X) tienen medias nulas.

Sustituyendo estos resultados en (1.57) se obtiene

$$\hat{s}_{\text{MMSE}} = \frac{v_S}{v_S + v_R} x \tag{1.60}$$

Este resultado puede ser interpretado de una manera bastante intuitiva: cuando la varianza del ruido es mucho menor que la de la señal (Relación Señal a Ruido (SNR) alta, $v_S \gg v_R$) se tiene que $\hat{s}_{\text{MMSE}} \to x$, lo que tiene sentido ya que el efecto de la componente de ruido en este caso no es muy significativo; por el contrario, cuando la SNR es muy baja ($v_S \ll v_R$), la observación apenas aporta información acerca del valor de S en cada experimento, por lo que el estimador se queda con el valor medio de la componente de señal, $\hat{s}_{\text{MMSE}} \to 0$.

1.5.2 Caso con variables multidimensionales

En un caso general multidimensional, $\mathbf{S}$ y $\mathbf{X}$ pueden ser vectores aleatorios de dimensiones N y M, respectivamente, con distribución conjuntamente gaussiana

$$p_{\mathbf{S},\mathbf{X}}(\mathbf{s},\mathbf{x}) \sim G\left(\begin{bmatrix}\mathbf{m_S}\\ \mathbf{m_X}\end{bmatrix}, \begin{bmatrix}\mathbf{V_S} & \mathbf{V_{SX}}\\ \mathbf{V_{SX}}^T & \mathbf{V_X}\end{bmatrix}\right) \tag{1.61}$$

siendo $\mathbf{m_S}$ y $\mathbf{m_X}$ las medias de $\mathbf{S}$ y $\mathbf{X}$, respectivamente, $\mathbf{V_S}$ y $\mathbf{V_X}$ las matrices de covarianzas de $\mathbf{S}$ y $\mathbf{X}$, respectivamente, y $\mathbf{V_{SX}}$ la matriz de covarianzas cruzadas de $\mathbf{S}$ y $\mathbf{X}$, y, de tal modo que

$$\mathbf{V_S} = \mathbb{E}\{(\mathbf{S}-\mathbf{m_S})(\mathbf{S}-\mathbf{m_S})^T\} \tag{1.62}$$

$$\mathbf{V_X} = \mathbb{E}\{(\mathbf{X}-\mathbf{m_X})(\mathbf{X}-\mathbf{m_X})^T\} \tag{1.63}$$

$$\mathbf{V_{SX}} = \mathbb{E}\{(\mathbf{S}-\mathbf{m_S})(\mathbf{X}-\mathbf{m_X})^T\} \tag{1.64}$$

El cálculo de la distribución a posteriori de $\mathbf{S}$ dado $\mathbf{X}$ es algo más complejo que en el caso unidimensional pero sigue un procedimiento similar, que omitiremos aquí. Puede demostrarse que la distribución a posteriori es gaussiana de media

$$\mathbf{m_{S|X}} = \mathbf{m_S} + \mathbf{V_{SX}}\mathbf{V_X}^{-1}(\mathbf{x}-\mathbf{m_X}) \tag{1.65}$$

y matriz de covarianzas

$$\mathbf{V_{S|X}} = \mathbf{V_S} - \mathbf{V_{SX}}\mathbf{V_X}^{-1}\mathbf{V_{SX}}^T \tag{1.66}$$

Dado que el estimador MMSE de $\mathbf{S}$ dado $\mathbf{X}$ es precisamente la media condicional, podemos escribir

$$\hat{\mathbf{s}}_{\text{MMSE}} = \mathbf{m_S} + \mathbf{V_{SX}}\mathbf{V_X}^{-1}(\mathbf{x}-\mathbf{m_X}) \tag{1.67}$$

La expresión del estimador se simplifica cuando $\mathbf{S}$ y $\mathbf{X}$ tienen medias nulas, resultando

$$\hat{\mathbf{s}}_{\text{MMSE}} = \mathbf{m_{S|X}} = \mathbf{V_{SX}}\mathbf{V_X}^{-1}\mathbf{x} \tag{1.68}$$

Partiendo de (1.68) pueden obtenerse diversos casos particulares de interés en aplicaciones prácticas del procesado de señales. Algunos de ellos se analizan en el Apéndice 1.8.1.

1.6 Estimación con restricciones

1.6.1 Principios generales

En ocasiones, puede resultar útil imponer una forma paramétrica determinada al estimador, $\hat{S} = f_{\mathbf{w}}(\mathbf{X})$, donde $\mathbf{w}$ es un vector que contiene todos los parámetros de la función. Por ejemplo, en un caso con dos observaciones $\mathbf{X} = [X_1, X_2]^T$, podría ser un requisito de diseño el restringir la búsqueda del estimador a la familia de estimadores cuadráticos de la forma $\hat{S} = w_0 + w_1 X_1^2 + w_2 X_2^2$. En estos casos, la tarea de diseño del estimador consiste en encontrar el vector óptimo de parámetros $\mathbf{w}^*$ que proporciona un mínimo coste medio sujeto a la restricción impuesta en la arquitectura del estimador:

$$\begin{aligned}\mathbf{w}^* &= \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbb{E}\{c(S,\hat{S})\} = \underset{\mathbf{w}}{\operatorname{argmin}}\ \mathbb{E}\{c(S, f_{\mathbf{w}}(\mathbf{X}))\} \\ &= \underset{\mathbf{w}}{\operatorname{argmin}} \int_{\mathbf{x}}\int_{s} c(s, f_{\mathbf{w}}(\mathbf{x})) p_{S,\mathbf{X}}(s,\mathbf{x}) ds d\mathbf{x} \end{aligned} \tag{1.69}$$

Puede entenderse fácilmente que la imposición de restricciones en la forma analítica del estimador hace que el estimador resultante incurra en un coste medio mayor que el que se obtendría utilizando el estimador bayesiano asociado a la misma función de coste[4]. No obstante, pueden existir razones de tipo práctico que hagan preferible el uso del primero, por ejemplo por simplicidad en el diseño o aplicación del estimador. Un ejemplo de esto lo tendremos en la Sección 1.6.2, dedicada al estudio de estimadores lineales de mínimo error cuadrático medio.

Example 1.18 (Cálculo de un estimador con restricciones).

Continuando el ejemplo 1.6, se desea calcular el estimador de mínimo error cuadrático medio que tenga la forma $\hat{s} = wx^2$. Partiendo del coste medio dado la observación calculado en (1.16), se puede obtener las expresión del coste medio global como

$$\begin{aligned}\mathbb{E}\{c(S,\hat{S})\} &= \int_{\mathbf{x}} \mathbb{E}\{c(S,\hat{s})|X = x\}\ p_{\mathbf{X}}(\mathbf{x})\mathbf{dx} \\ &= \int_{\mathbf{x}} \left(\frac{1}{3}x^2 - \hat{s}x + \hat{s}^2\right) p_{\mathbf{X}}(\mathbf{x})\mathbf{dx}\end{aligned} \tag{1.70}$$

Forzando $\hat{s} = wx^2$ y teniendo en cuenta que $p_{\mathbf{X}}(\mathbf{x}) = \mathbf{1}$ para $0 < x < 1$, se obtiene el coste medio global en función de w

$$\mathbb{E}\{c(S, w\mathbf{X}^2)\} = \int_{\mathbf{x}} \left(\frac{1}{3}x^2 - wx^3 + w^2x^4\right) d\mathbf{x} \tag{1.71}$$

$$= \frac{1}{9} - \frac{1}{4}w + \frac{1}{5}w^2 \tag{1.72}$$

[4] La única excepción a esta regla consiste precisamente en el caso en el que las restricciones impuestas permiten obtener el estimador óptimo o, dicho de otro modo, cuando el estimador bayesiano presenta una forma analítica compatible con las restricciones impuestas.

El valor w^* que optimiza (1.72) puede calcularse derivando respecto de w e igualando a cero la expresión obtenida:

$$\frac{d}{d\hat{w}}\mathbb{E}\{c(S, w\mathbf{X}^2)\}\Big|_{w=w^*} = -\frac{1}{4} + \frac{2}{5}w^* = 0, \tag{1.73}$$

$$w^* = \frac{5}{8}, \tag{1.74}$$

y por lo tanto el estimador buscado es: $\hat{s} = \frac{5}{8}x^2$.

1.6.2 Estimación lineal de mínimo error cuadrático medio

En esta sección nos centraremos en el estudio de estimadores de variable aleatoria que obtienen su salida como combinación lineal de los valores de las observaciones, utilizando la minimización del coste cuadrático medio como criterio de diseño. Por lo tanto, consideraremos exclusivamente estimadores que calculan su salida como

$$\hat{S} = w_0 + w_1 X_1 + \cdots + w_N X_N \tag{1.75}$$

donde N denota el número de variables observables disponibles, $\{X_i\}_{i=1}^N$, y $\{w_i\}_{i=0}^N$ son los pesos que caracterizan al estimador. En este contexto, es habitual referirse al término independiente de la expresión anterior, w_0, como término de sesgo. Por simplicidad analítica, resulta más cómodo introducir la siguiente notación matricial:

$$\hat{S} = w_0 + \mathbf{w}^T\mathbf{X} = \mathbf{w}_\mathrm{e}^T\mathbf{X}_\mathrm{e} \tag{1.76}$$

donde $\mathbf{w} = [w_1, \ldots, w_N]^T$ y $\mathbf{X} = [X_1, \ldots, X_N]^T$ son los vectores (columna) de parámetros y de observaciones, respectivamente, y $\mathbf{w}_\mathrm{e} = [w_0, \mathbf{w}^T]^T$ y $\mathbf{X}_\mathrm{e} = [1, \mathbf{X}^T]^T$ son versiones extendidas de dichos vectores.

Puede entenderse que, al imponer una restricción en la forma analítica que implementa el estimador, los estimadores lineales obtendrán, en general, prestaciones inferiores al estimador bayesiano óptimo. No obstante, el interés de los estimadores lineales está justificado por su mayor simplicidad y facilidad de diseño. Como veremos, para el cálculo del estimador lineal de mínimo error cuadrático medio, será suficiente conocer los momentos estadísticos de primer y segundo orden (medias y covarianzas) asociados a las variables observables y la variable a estimar.

Por otro lado, el empleo de estimadores lineales está plenamente justificado en ciertas circunstancias, por ejemplo al tratar con variables con distribuciones gaussianas, ya que, como vimos en la sección anterior, en dicho caso el estimador bayesiano de mínimo error cuadrático medio tiene arquitectura lineal.

Minimización del error cuadrático medio

Como ya se ha comentado, consideraremos como criterio de diseño el coste cuadrático, $c(e) = (s - \hat{s})^2$, por lo que el vector de pesos óptimo será aquel que minimice el valor medio de dicha función de coste:

$$\mathbf{w}_{\mathrm{e}}^* = \underset{\mathbf{w}_{\mathrm{e}}}{\operatorname{argmin}}\ \mathbb{E}\{(S - \hat{S})^2\} = \underset{\mathbf{w}_{\mathrm{e}}}{\operatorname{argmin}}\ \mathbb{E}\{(S - \mathbf{w}_{\mathrm{e}}^T \mathbf{X}_{\mathrm{e}})^2\} \tag{1.77}$$

y nos referiremos al estimador lineal asociado a dicho vector óptimo de pesos como $\hat{S}_{\mathrm{LMSE}}$:

$$\hat{S}_{\mathrm{LMSE}} = \mathbf{w}_{\mathrm{e}}^{*T} \mathbf{X}_{\mathrm{e}}$$

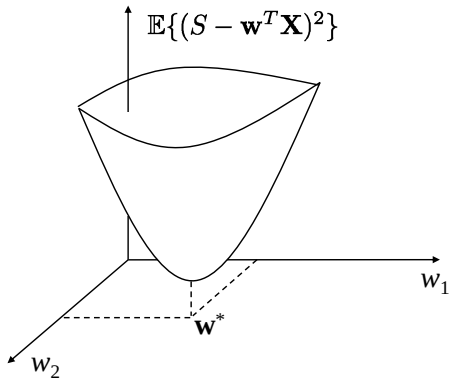


Fig. 1.7. Superficie de error cuadrático medio de un estimador lineal de variable aleatoria como función de los pesos del estimador.

La Figura 1.7 representa la superficie de error en un caso con dos observaciones. Al ser la función a minimizar cuadrática en los pesos (argumento de la minimización), la superficie de error tendrá forma de un paraboloide de N dimensiones. Además, dado que el coste medio es no negativo, queda garantizado que la función es convexa, y su mínimo puede localizarse igualando a $\mathbf{0}$ el gradiente del coste medio con respecto del vector de pesos[5]:

$$\begin{aligned} \nabla_{\mathbf{w}_{\mathrm{e}}} \mathbb{E}\{(S - \hat{S})^2\}\Big|_{\mathbf{w}_{\mathrm{e}} = \mathbf{w}_{\mathrm{e}}^*} &= -2\mathbb{E}\{(S - \mathbf{w}_{\mathrm{e}}^T \mathbf{X}_{\mathrm{e}})\mathbf{X}_{\mathrm{e}}\}\big|_{\mathbf{w}_{\mathrm{e}} = \mathbf{w}_{\mathrm{e}}^*} = \\ &= -2\mathbb{E}\{(S - \mathbf{w}_{\mathrm{e}}^{*T} \mathbf{X}_{\mathrm{e}})\mathbf{X}_{\mathrm{e}}\} = \mathbf{0} \end{aligned} \tag{1.78}$$

[5] El gradiente de una función escalar $f(\mathbf{w})$ con respecto del vector $\mathbf{w}$ se define como un vector formado por las derivadas de la función con respecto de cada una de las componentes de $\mathbf{w}$: $\nabla_{\mathbf{w}} f(\mathbf{w}) = \left[\frac{\partial f}{\partial w_1}, \ldots \frac{\partial f}{\partial w_N}\right]^T$.

La segunda línea de la expresión anterior define las condiciones que debe cumplir el vector de pesos óptimo. Nótese que dicha ecuación constituye, en realidad, un sistema de $N+1$ ecuaciones (tantas como dimensiones tiene $\mathbf{X}_\mathrm{e}$) con $N+1$ incógnitas (las componentes de $\mathbf{w}_\mathrm{e}^*$).

Para encontrar el vector óptimo de pesos, resulta conveniente reescribir la última línea de (1.78) como

$$\mathbb{E}\{S\mathbf{X}_\mathrm{e}\} = \mathbb{E}\{\mathbf{X}_\mathrm{e}(\mathbf{X}_\mathrm{e}^T\mathbf{w}_\mathrm{e}^*)\} \tag{1.79}$$

Definiendo el vector de correlación cruzada

$$\mathbf{r}_{S\mathbf{X}_e} = \mathbb{E}\{S\mathbf{X}_e\} \tag{1.80}$$

y la matrix de correlación

$$\mathbf{R}_{\mathbf{X}_e} = \mathbb{E}\{\mathbf{X}_e\mathbf{X}_e^T\} \tag{1.81}$$

(que es una matriz simétrica) la ec. (1.79) se puede escribir como

$$\mathbf{r}_{S\mathbf{X}_e} = \mathbf{R}_{\mathbf{X}_e}\mathbf{w}_\mathrm{e}^* \tag{1.82}$$

De donde resulta el vector de coeficientes buscado:

$$\boxed{\mathbf{w}_\mathrm{e}^* = \mathbf{R}_{\mathbf{X}_e}^{-1}\mathbf{r}_{S\mathbf{X}_e}} \tag{1.83}$$

Propiedades del estimador lineal óptimo

La ecuación (1.82) resuelve el problema del cálculo de los pesos del estimador $\hat{S}_\mathrm{LMSE}$. Pero resulta interesante volver sobre la ecuación vectorial (1.78) para analizar algunas de sus propiedades. Obsérvese que el término entre paréntesis en esta ecuación constituye el error de estimación

$$E^* = S - \mathbf{w}_\mathrm{e}^{*T}\mathbf{X}_\mathrm{e} \tag{1.84}$$

de modo que podemos reescribir (1.78) como

$$\mathbb{E}\{E^*\mathbf{X}_\mathrm{e})\} = \mathbf{0} \tag{1.85}$$

Tomando, por un lado, la primera componente de esta ecuación (teniendo en cuenta que $X_{\mathrm{e},1} = 1$, y el resto por otro, se obtienen dos propiedades fundamentales del estimador lineal de mínimo error cuadrático medio:

Propiedad 1: El error tiene media nula:

$$\mathbb{E}\{E^*\} = \mathbf{0} \tag{1.86}$$

Cuando un estimador tiene esta propiedad se dice que es **insesgado**. Volveremos sobre esta propiedad en la sec. 1.7.

Propiedad 2 (Principio de Ortogonalidad): el error es estadísticamente ortogonal a las observaciones:

$$\mathbb{E}\{E^*\mathbf{X}\} = \mathbf{0} \tag{1.87}$$

Expresión alternativa del estimador

Expandiendo las ecs. (1.86) y (1.87), podemos obtener las siguientes fórmulas explícitas para los coeficientes w_0^* y $\mathbf{w}^*$ del estimador.

$$w_0^* = m_S - \mathbf{w}^{*T}\mathbf{m}_\mathbf{x} \tag{1.88}$$

$$\mathbf{w}^* = \mathbf{V}_\mathbf{X}^{-1}\,\mathbf{v}_{S,\mathbf{X}} \tag{1.89}$$

Se puede observar que el papel del término de sesgo w_0 consiste en compensar las diferencias entre las medias de la variable a estimar y las observaciones. Por lo tanto, cuando todas las variables involucradas tengan medias nulas, se tendrá que $w_0^* = 0$. En contraposición al papel de w_0, podemos afirmar que el vector de pesos $\mathbf{w}$ permite minimizar el error cuadrático medio de las fluctuaciones de S alrededor de su media, explotando para ello la relación estadística existente entre S y $\mathbf{X}$.

Dedicaremos este apartado a obtener las expresiones (1.88) y (1.89). La primera es una consecuencia directa de (1.86) que puede desarrollarse como

$$m_S - \mathbf{w}^{*T}\mathbf{m}_\mathbf{x} - w_0^* = 0 \tag{1.90}$$

despejando w_0^* se llega a (1.88).

Buscaremos ahora una expresión para $\mathbf{w}^*$. De (1.87) resulta

$$\mathbb{E}\{(S - \mathbf{w}^{*T}\mathbf{X} - w_0^*)\mathbf{X}\} = \mathbf{0} \tag{1.91}$$

que puede reescribirse como

$$\begin{aligned}\mathbb{E}\{S\mathbf{X}\} &= \mathbb{E}\{(\mathbf{w}^{*T}\mathbf{X} + w_0^*)\mathbf{X}\} \\ &= \mathbb{E}\{\mathbf{X}(\mathbf{X}^T\mathbf{w}^*)\} + w_0^*\mathbb{E}\{\mathbf{X}\} \\ &= \mathbb{E}\{\mathbf{X}\mathbf{X}^T\}\mathbf{w}^* + w_0^*\mathbf{m}_\mathbf{X}\end{aligned} \tag{1.92}$$

Recurriendo ahora a las expresiones que relacionan la correlación y la covarianza de dos variables:

$$\mathbb{E}\{S\mathbf{X}\} = \mathbf{v}_{S,\mathbf{X}} + m_S\mathbf{m}_\mathbf{X} \tag{1.93}$$

$$\mathbb{E}\{\mathbf{X}\mathbf{X}^T\} = \mathbf{V}_\mathbf{X} + \mathbf{m}_\mathbf{X}\mathbf{m}_\mathbf{X}^T \tag{1.94}$$

la ec. (1.92) se convierte en

$$\begin{aligned}\mathbf{v}_{S,\mathbf{X}} &= \mathbf{V}_\mathbf{X}\mathbf{w}^* - \mathbf{m}_\mathbf{X}\mathbf{m}_\mathbf{X}^T\mathbf{w}^* + w_0^*\mathbf{m}_\mathbf{X} - m_S\mathbf{m}_\mathbf{X} \\ &= \mathbf{V}_\mathbf{X}\mathbf{w}^* + \mathbf{m}_\mathbf{X}(w_0^* - \mathbf{m}_\mathbf{X}^T\mathbf{w}^* - m_S) \\ &= \mathbf{V}_\mathbf{X}\mathbf{w}^*\end{aligned} \tag{1.95}$$

donde, en la última igualdad, hemos aplicado (1.88). Por tanto, despejando $\mathbf{w}^*$, se obtiene (1.89)

Estimación lineal y estimación gaussiana

Aplicando (1.89) y (1.88) sobre (1.76), el estimador lineal de mínimo error cuadrático medio puede escribirse como

$$\hat{s}_{\text{LMSE}} = (\mathbf{w}^*)^T\mathbf{x} + w_0^* = m_S + \mathbf{v}_{S,\mathbf{X}}^T\mathbf{V}_{\mathbf{X}}^{-1}(\mathbf{x} - \mathbf{m}_{\mathbf{X}}) \tag{1.96}$$

Resulta interesante comprobar que esta expresión coincide con (1.65) para $\mathbf{S}$ unidimesional. Esto no es sorprendente: dado que el estimador MMSE sin restricciones en el caso gaussiano es lineal, el mejor estimador lineal debe coincidir con el obtenido para el caso gaussiano.

Obsérvese, por último, que (1.89) asume que $\mathbf{V}_{\mathbf{X}}$ es una matriz no singular. La invertibilidad de $\mathbf{V}_{\mathbf{X}}$ implica que ninguna componente de $\mathbf{X}$ puede obtenerse como combinación lineal del resto de componentes. Cuando esto no es así, puede comprobarse que la solución al problema de minimización no es única, y por lo tanto conviene eliminar las variables redundantes antes de proceder al diseño del estimador.

Error cuadrático medio mínimo

Calcularemos aquí el error cuadrático medio asociado al estimador lineal de mínimo error cuadrático medio, $\hat{S}_{\text{LMSE}}$. Como se comentó al inicio de esta sección, el error cuadrático medio obtenido será, en general, superior al que obtendría el estimador bayesiano de mínimo error cuadrático medio ($\hat{S}_{\text{MMSE}}$) para el mismo problema, salvo cuando este último estimador tenga precisamente estructura lineal.

Para calcular el error cuadrático medio no tenemos más que desarrollar la expresión del coste medio, particularizándola para $\hat{S}_{\text{LMSE}}$, dejando el resultado en función de las esperanzas matemáticas de las variables aleatorias:

$$\begin{aligned}\mathbb{E}\{(S - \hat{S}_{\text{LMSE}})^2\} &= \mathbb{E}\{E^*(S - w_0^* - \mathbf{w}^{*T}\mathbf{X})\} \\ &= \mathbb{E}\{E^*S\} - w_0^*\mathbb{E}\{E^*\} - \mathbf{w}^{*T}\mathbb{E}\{\mathbf{X}E^*\} \\ &= \mathbb{E}\{E^*S\}\end{aligned} \tag{1.97}$$

donde, en la última igualdad, hemos aplicado las dos propiedades del estimador de mínimo error cuadrático medio obtenidas en (1.86) y (1.87). Desarrollando de nuevo el término de error, E^*, resulta

$$\begin{aligned}\mathbb{E}\{(S - \hat{S}_{\text{LMSE}})^2\} &= \mathbb{E}\{S(S - w_0^* - \mathbf{w}^{*T}\mathbf{X})\} \\ &= \mathbb{E}\{S^2\} - w_0^*m_S - \mathbf{w}^{*T}(\mathbf{v}_{S\mathbf{X}} + m_S\mathbf{m}_{\mathbf{X}})\} \\ &= \mathbb{E}\{S^2\} - m_S(w_0^* + \mathbf{w}^{*T}\mathbf{m}_{\mathbf{X}}) - \mathbf{w}^{*T}\mathbf{v}_{S\mathbf{X}} \\ &= v_S - \mathbf{w}^{*T}\mathbf{v}_{S\mathbf{X}}\end{aligned} \tag{1.98}$$

Exercise 1.19 (Estimación lineal de mínimo error cuadrático medio). Se desea construir un estimador lineal de mínimo error cuadrático medio que permita estimar la variable aleatoria S a partir de las variables aleatorias X_1 y X_2. Sabiendo que

$$\begin{array}{lll} \mathbb{E}\{S\} = 1/2 & \mathbb{E}\{X_1\} = 1 & \mathbb{E}\{X_2\} = 0 \\ \mathbb{E}\{S^2\} = 4 & \mathbb{E}\{X_1^2\} = 3/2 & \mathbb{E}\{X_2^2\} = 2 \\ \mathbb{E}\{SX_1\} = 1 & \mathbb{E}\{SX_2\} = 2 & \mathbb{E}\{X_1X_2\} = 1/2 \end{array}$$

obténganse los pesos del estimador buscado y calcúlese su error cuadrático medio. Calcúlese el valor estimado para el siguiente vector de observaciones: $[X_1, X_2] = [3, 1]$.

Example 1.20 (Extensión al caso multidimensional). A lo largo de la discusión teórica previa se consideró en exclusiva el caso en que la variable a estimar tiene carácter unidimensional. Cuando se desea construir el estimador lineal de mínimo error cuadrático medio de un vector aleatorio $\mathbf{S}$, el problema puede formularse como

$$\hat{\mathbf{S}} = \mathbf{w}_0 + \mathbf{W}^T\mathbf{X}$$

donde $\mathbf{W}$ es ahora una matriz que contiene tantas columnas como variables a estimar, y tantas filas como observaciones disponibles, mientras que $\mathbf{w}_0$ es un vector columna de términos de sesgo.

La solución a este problema puede obtenerse como extensión directa del caso unidimensional, y está caracterizada por

$$\mathbf{W}^* = \mathbf{V}_{\mathbf{X}}^{-1}\,\mathbf{V}_{\mathbf{S},\mathbf{X}}^T$$

$$\mathbf{w}_0^* = \mathbb{E}\{\mathbf{S}\} - \mathbf{W}^{*T}\mathbb{E}\{\mathbf{X}\}$$

siendo $\mathbf{V}_{\mathbf{S},\mathbf{X}}$ la matriz de covarianzas cruzadas entre los vectores aleatorios $\mathbf{S}$ y $\mathbf{X}$. Puede comprobarse que, como cabría esperar, al calcular el estimador $\hat{\mathbf{S}}_{\text{LMSE}}$ para este caso, resulta la misma expresión que obtuvimos en (1.65) para el caso gaussiano sin restricciones

La estimación lineal de mínimo error cuadrático medio y el Principio de Ortogonalidad presentan algunas analogías con la aproximación lineal de vectores en espacios vectoriales. El lector interesado puede acudir al apéndice 1.8.2.

1.7 Caracterización de estimadores

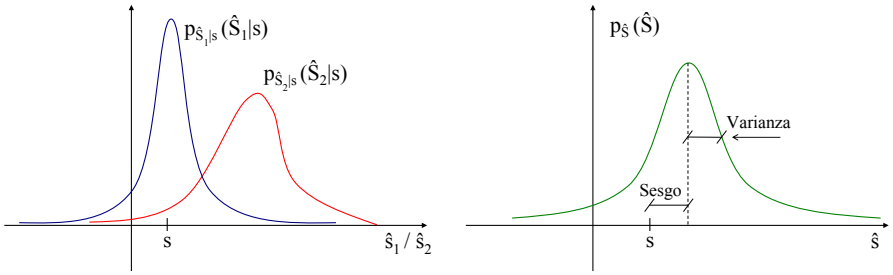
A lo largo de este capítulo hemos presentado diversos métodos de estimación, comprobando que para un mismo escenario de aplicación es posible diseñar diferentes estimadores no triviales. Por tanto, surge la necesidad de establecer criterios que permitan una comparación objetiva entre estimadores. Una primera posibilidad para evaluar las prestaciones de un estimador es evaluar su coste medio para una determinada función de coste. Queda claro, no obstante, que ningún estimador ofrecerá un menor coste medio que el estimador bayesiano asociado a dicha función de coste.

En esta sección analizamos otras medidas que permiten obtener una primera aproximación acerca de las propiedades de un estimador. En concreto, introduciremos los conceptos de sesgo y de varianza, que dan idea del error sistemático y de la

dispersión de las estimaciones frente a un valor medio (recuérdese el carácter aleatorio de $\hat{S}$). Por simplicidad, comenzaremos considerando el caso de estimación de parámetro determinista, para pasar posteriormente a extender estos conceptos a la estimación de variable aleatoria.

1.7.1 Sesgo y varianza de estimadores de parámetros deterministas

Una caracterización completa del comportamiento de un estimador de parámetro determinista la proporciona la densidad de probabilidad del estimador para cada posible valor del parámetro a estimar, es decir, $p_{\hat{S}|s}(\hat{s}|s)$. Nótese, que al ser el estimador una función de las observaciones, $\hat{S} = f(\mathbf{X})$, es posible obtener dicha densidad de probabilidad a partir de la de $\mathbf{X}$ (dado s), aplicando el cambio de variable aleatoria correspondiente.



(a) (b)

Fig. 1.8. Sesgo y varianza de estimadores de parámetro determinista. La figura de la izquierda muestra las densidades de probabilidad asociadas a dos estimadores diferentes, mientras que la figura de la derecha ilustra el significado físico del sesgo y la varianza de un estimador.

La Figura 1.8(a) muestra la distribución de probabilidad que se obtendría con dos estimadores diferentes, $\hat{S}_1 = f_1(\mathbf{X})$ y $\hat{S}_2 = f_2(\mathbf{X})$, y sugiere que, en este caso concreto, el empleo del primero de los estimadores será en general más beneficioso, ya que la probabilidad de estimar valores cercanos al valor real de s es mucho mayor que si usáramos $\hat{S}_2$. Nótese que esto no implica que en cada aplicación concreta de los estimadores $\hat{S}_1$ obtenga menor error de estimación.

Para disponer de una caracterización más cómoda de los estimadores resulta útil resumir algunas de las propiedades más relevantes de $p_{\hat{S}}(\hat{s})$, y la forma más evidente de hacer esto es mediante la media y la varianza de dicha distribución. En realidad, más que interesarnos la media de la distribución nos interesa cómo de alejada está dicha media del valor real de s, siendo ésta la definición del *sesgo* de un estimador:

$$\text{Sesgo}(\hat{S}) = \mathbb{E}\{s - \hat{S}\} = s - \mathbb{E}\{\hat{S}\} \tag{1.99}$$

$$\text{Varianza}(\hat{S}) = \mathbb{E}\{(\hat{S} - \mathbb{E}\{\hat{S}\})^2\} = \mathbb{E}\{\hat{S}^2\} - \mathbb{E}^2\{\hat{S}\} \tag{1.100}$$

Cabe mencionar que, cuando s es un parámetro determinista, la varianza del estimador coincide con la de el error de estimación, ya que Varianza$(s - \hat{S}) =$ Varianza$(\hat{S})$.

Es importante resaltar que, en el caso de estimación de parámetro determinista, el sesgo y la varianza son funciones de la variable que se desea estimar (s). Nótese que todas las esperanzas matemáticas de las expresiones anteriores pueden ser calculadas tanto a partir de la función de densidad de probabilidad de $\mathbf{X}$ como de la de $\hat{S}$. Nuevamente, es posible denotar la dependencia de dichas densidades con s a la hora de calcular las esperanzas matemáticas, por ejemplo,

$$\begin{aligned}\mathbb{E}\{\hat{S}^2\} = \mathbb{E}\{\hat{S}^2|s\} &= \int \hat{s}^2 \; p_{\hat{S}}(\hat{s}) d\hat{s} = \int \hat{s}^2 \; p_{\hat{S}|s}(\hat{s}|s) d\hat{s} \\ &= \int f^2(\mathbf{x}) \; p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int f^2(\mathbf{x}) \; p_{\mathbf{X}|s}(\mathbf{x}|s) d\mathbf{x}\end{aligned} \tag{1.101}$$

La Figura 1.8(b) ilustra el significado físico del sesgo y la varianza de un estimador. Como puede verse, el sesgo tiene significado de error sistemático, es decir, sería la media de los errores que se obtendrían si aplicásemos el estimador un número infinito de veces con distintas observaciones. A los estimadores que tienen sesgo nulo se les denomina *estimadores insesgados*. Por otro lado, la varianza da idea de cómo de concentrada está la probabilidad del estimador en torno a su media y, por lo tanto, está relacionada con la dispersión de los valores que se observarían al aplicar el estimador sobre distintas observaciones.

En el caso particular de estimadores que operan sobre un número l de observaciones de una variable aleatoria, por ejemplo al estimar la media o la varianza de una distribución de tipo Gauss a partir de l observaciones de la misma (Ejemplo 1.14), una propiedad deseable es que la varianza del estimador decrezca conforme aumenta el número de observaciones, i.e., Varianza$(\hat{S}) \to 0$ cuando $l \to \infty$. A los estimadores que disfrutan de esta propiedad se los conoce como estimadores *consistentes en varianza*.

Por último, es posible relacionar el sesgo y la varianza con el coste cuadrático medio del estimador:

$$\begin{aligned}\mathbb{E}\{(s - \hat{S})^2\} &= \text{Varianza}\{s - \hat{S}\} + \mathbb{E}^2\{s - \hat{S}\} \\ &= \text{Varianza}(\hat{S}) + [\text{Sesgo}(\hat{S})]^2\end{aligned} \tag{1.102}$$

Example 1.21 (Cálculo del sesgo y la varianza del estimador muestral de la media de una distribución). El estimador muestral de la media m de una variable aleatoria X a partir de l observaciones independientes de la misma, $\{X^{(k)}\}_{k=1}^{l}$, se define como

$$\hat{M} = \frac{1}{l} \sum_{k=1}^{l} X^{(k)} \tag{1.103}$$

Podemos calcular el sesgo y la varianza de dicho estimador de manera sencilla como

$$\text{Sesgo}(\hat{M}) = m - \mathbb{E}\{\hat{M}\} = m - \frac{1}{l}\sum_{k=1}^{l}\mathbb{E}\{X^{(k)}\} = 0$$

$$\text{Varianza}(\hat{M}) = \text{Varianza}\left(\frac{1}{l}\sum_{k=1}^{l}X^{(k)}\right) = \frac{1}{l^2}\sum_{k=1}^{l}\text{Varianza}(X^{(k)}) = \frac{v}{l}$$

En el cálculo de la varianza del estimador, se ha utilizado v para denotar la media de la variable aleatoria X, y se ha utilizado además el hecho de que las observaciones son independientes. A la vista de los resultados, puede comprobarse que el estimador muestral de la media es insesgado y consistente en varianza.

Example 1.22 (Cálculo del sesgo del estimador muestral de la varianza de una distribución). El estimador muestral de la varianza v de una variable aleatoria X a partir de l observaciones independientes de la misma, $\{X^{(k)}\}_{k=1}^{l}$, se define como

$$\hat{V} = \frac{1}{l}\sum_{k=1}^{l}(X^{(k)} - \hat{M})^2 \tag{1.104}$$

donde $\hat{M}$ es el estimador muestral de la media dado por (1.103).

Podemos calcular el sesgo de dicho estimador como

$$\begin{aligned}
\text{Sesgo}(\hat{V}) &= v - \mathbb{E}\{\hat{V}\} = v - \frac{1}{l}\sum_{k=1}^{l}\mathbb{E}\{(X^{(k)} - \hat{M})^2\} \\
&= v - \frac{1}{l}\sum_{k=1}^{l}\left[\mathbb{E}\{{X^{(k)}}^2\} + \mathbb{E}\{\hat{M}^2\} - 2\mathbb{E}\{X^{(k)}\hat{M}\}\right] \\
&= v - \frac{1}{l}\sum_{k=1}^{l}\left[v + m^2 + \text{Varianza}(\hat{M}) + \mathbb{E}^2\{\hat{M}\} - 2\frac{1}{l}\mathbb{E}\{X^{(k)}\sum_{k=1}^{l}X^{(k)}\}\right] \\
&= v - \frac{1}{l}\sum_{k=1}^{l}\left[v + m^2 + \frac{v}{l} + m^2 - 2\frac{1}{l}\left[\mathbb{E}\{{X^{(k)}}^2\} + (l-1)\mathbb{E}^2\{X^{(k)}\}\right]\right] \\
&= v - \frac{1}{l}\sum_{k=1}^{l}\left[v + 2m^2 + \frac{v}{l} - 2\frac{1}{l}\left[v + m^2 + (l-1)m^2\right]\right] \\
&= v - \frac{(l-1)}{l}v
\end{aligned}$$

Dado que $\mathbb{E}\{\hat{V}\} = \frac{l-1}{l}v \neq v$, el estimador muestral de la varianza es sesgado, si bien es asintóticamente insesgado, ya que según crece el número de observaciones el sesgo tiende a cero.

Exercise 1.23 (Estimador insesgado de la varianza). Se desea corregir el estimador muestral de la varianza, de modo que el nuevo estimador sea insesgado independientemente del número de observaciones. Para ello se decide utilizar una versión escalada del estimador:

$$\hat{V}_{\text{ins}} = c\,\hat{V}$$

donde $\hat{V}$ es el estimador muestral de la varianza, $\hat{V}_{\text{ins}}$ es el estimador buscado, y c es una constante a determinar. Obtenga el valor de la constante c que hace que $\hat{V}_{\text{ins}}$ sea insesgado. Demuestre que la varianza del estimador insesgado es mayor que la que se obtendría al utilizar el estimador muestral. Este resultado ilustra el importante compromiso entre sesgo y varianza que aparece con frecuencia en problemas de estimación: resulta posible disminuir la varianza (el sesgo) de un estimador a costa de un incremento de su sesgo (varianza).

1.7.2 Sesgo y varianza de estimadores de variables aleatorias

La extensión de los conceptos de sesgo y varianza para el caso de estimación de variable aleatoria resulta inmediata. De hecho, y de forma análoga al caso determinista, sería posible utilizar directamente la distribución $p_{\hat{S}|S}(\hat{s}|s)$ para obtener información acerca de la bondad de un estimador para cada posible valor de la variable aleatoria. Sin embargo, al aplicar repetidas veces un estimador de variable aleatoria, el valor s de la variable a estimar cambia de experimento a experimento y, por este motivo, resulta necesario obtener también la esperanza matemática con respecto de S para tener una idea precisa acerca del error sistemático que se obtiene al aplicar el estimador.

Por lo tanto, en el caso de estimación de variable aleatoria definimos el sesgo y la varianza como

$$\text{Sesgo}(\hat{S}) = \mathbb{E}\{S - \hat{S}\} = \mathbb{E}\{S\} - \mathbb{E}\{\hat{S}\} \tag{1.105}$$

$$\text{Varianza}(\hat{S}) = \mathbb{E}\{(\hat{S} - \mathbb{E}\{\hat{S}\})^2\} = \mathbb{E}\{\hat{S}^2\} - \mathbb{E}^2\{\hat{S}\} \tag{1.106}$$

En este caso, es posible llevar a cabo una descomposición del error cuadrático medio similar a la utilizada para el caso determinista:

$$\begin{aligned} \mathbb{E}\{(S - \hat{S})^2\} &= \text{Varianza}\{S - \hat{S}\} + \mathbb{E}^2\{S - \hat{S}\} \\ &= \text{Varianza}(E) + [\text{Sesgo}(\hat{S})]^2 \end{aligned} \tag{1.107}$$

donde E es error de estimación en que incurre el estimador $\hat{S}$. Nótese que, al contrario de lo que ocurría en el caso determinista, cuando la variable a estimar S es aleatoria la varianza del error no será en general igual a la varianza del estimador.

Conviene por último resaltar que el cálculo de las esperanzas matemáticas anteriores que involucran a S y $\hat{S}$ puede realizarse utilizando la distribución conjunta de dichas dos variables o, alternativamente, la distribución conjunta de S y $\mathbf{X}$, haciendo uso de la relación determinista que existe entre $\hat{S}$ y $\mathbf{X}$. Así, por ejemplo,

$$\begin{aligned}\mathbb{E}\{(S-\hat{S})^2\} &= \int_{(s)}\int_{(\hat{s})}(s-\hat{s})^2 p_{S,\hat{S}}(s,\hat{s})\; ds\; d\hat{s} \\ &= \int_{(s)}\int_{(\mathbf{x})}(s-f(\mathbf{x}))^2 p_{S,\mathbf{X}}(s,\mathbf{x})\; ds\; d\mathbf{x}\end{aligned} \tag{1.108}$$

Mencionaremos, finalmente, dos propiedades de interés relativas al sesgo:

- El estimador de mínimo error cuadrático medio (sin restricciones) $\mathbb{E}\{\hat{S}_{\text{MMSE}}\}$ es siempre insesgado:

$$\begin{aligned}\mathbb{E}\{\hat{S}_{\text{MMSE}}\} &= \mathbb{E}\left\{\mathbb{E}\{S|\mathbf{X}\}\right\} \\ &= \int \mathbb{E}\left\{S|\mathbf{X}=\mathbf{x}\right\} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\ &= \mathbb{E}\{S\}\end{aligned} \tag{1.109}$$

- Asimismo, el estimador lineal de mínimo error cuadrático medio también es insesgado. Esto es una consecuencia inmediata de la propiedad 1 en (1.86)

1.8 Apéndices

1.8.1 Casos particulares gaussianos

Partiendo de (1.65) pueden obtenerse estimadores MMSE para diferentes casos particulares de interés, que se analizan en los apartados siguientes.

Transformaciones lineales con ruido

Supongamos que la observación $\mathbf{X}$ está relacionada con $\mathbf{S}$ a través de la expresión.

$$\mathbf{X} = \mathbf{H}\mathbf{S} + \mathbf{R}$$

donde $\mathbf{H}$ es una matriz determinista conocida de dimensiones $M \times N$, y $\mathbf{R}$ es un vector gaussiano aleatorio de dimensiones $M \times 1$, independiente de $\mathbf{S}$. Las distribuciones de los vectores $\mathbf{S}$ y $\mathbf{R}$ son:

$$p_{\mathbf{S}}(\mathbf{s}) = G(\mathbf{0}, \mathbf{V}_{\mathbf{S}}) \qquad p_{\mathbf{R}}(\mathbf{r}) = G(\mathbf{0}, \mathbf{V}_{\mathbf{R}})$$

siendo $\mathbf{0}$ un vector columna con todas sus componentes iguales a 0.

De acuerdo con ésto, podemos comprobar que

$$\mathbb{E}\{\mathbf{X}\} = \mathbf{H}\mathbb{E}\{\mathbf{S}\} + \mathbb{E}\{\mathbf{R}\} = \mathbf{0} \tag{1.110}$$

Por tanto, $\mathbf{S}$ y $\mathbf{X}$ tienen media nula, y podemos aplicar la ecuación (1.68). Para ello, calcularemos $\mathbf{V}_{\mathbf{SX}}$ y $\mathbf{V}_{\mathbf{X}}$. En primer lugar,

$$\begin{aligned}\mathbf{V_{SX}} &= \mathbb{E}\{\mathbf{S}\mathbf{X}^T\} \\ &= \mathbb{E}\{\mathbf{S}(\mathbf{H}\mathbf{S}+\mathbf{R})^T\} \\ &= \mathbb{E}\{\mathbf{S}\mathbf{S}^T\}\mathbf{H}^T + \mathbb{E}\{\mathbf{S}\mathbf{R})^T \\ &= \mathbf{V_S}\mathbf{H}^T + \mathbb{E}\{\mathbf{S}\}\mathbb{E}\{\mathbf{R})^T \\ &= \mathbf{V_S}\mathbf{H}^T \end{aligned} \tag{1.111}$$

(donde, en la tercera igualdad, hemos hecho uso de la independencia de $\mathbf{S}$ y $\mathbf{R}$). Analogamente,

$$\begin{aligned}\mathbf{V_X} &= \mathbb{E}\{\mathbf{X}\mathbf{X}^T\} \\ &= \mathbb{E}\{(\mathbf{H}\mathbf{S}+\mathbf{R})(\mathbf{H}\mathbf{S}+\mathbf{R})^T\} \\ &= \mathbf{H}\mathbb{E}\{\mathbf{S}\mathbf{S}^T\}\mathbf{H}^T + \mathbb{E}\{\mathbf{R}\mathbf{R})^T \\ &= \mathbf{H}\mathbf{V_S}\mathbf{H}^T + \mathbf{V_R} \end{aligned} \tag{1.112}$$

(donde, de nuevo, en la tercera igualdad hemos hecho uso de la independencia de $\mathbf{S}$ y $\mathbf{R}$). Aplicando (1.111) y (1.112) en (1.68), resulta

$$\hat{\mathbf{s}}_{\text{MMSE}} = \mathbf{m}_{\mathbf{S}|\mathbf{X}} = \mathbf{V_S}\mathbf{H}^T(\mathbf{H}\mathbf{V_S}\mathbf{H}^T + \mathbf{V_R})^{-1}\mathbf{x} \tag{1.113}$$

Una expresión alternativa pero equivalente a la anterior puede obtenerse aplicando el denominado lema de inversión de la matriz, según el cual

$$(\mathbf{H}\mathbf{V_S}\mathbf{H}^T + \mathbf{V_R})^{-1} = \mathbf{V_R}^{-1} - \mathbf{V_R}^{-1}\mathbf{H}\left(\mathbf{H}^T\mathbf{V_R}^{-1}\mathbf{H} + \mathbf{V_S}^{-1}\right)\mathbf{H}^T{\mathbf{V_R}}^{-1} \tag{1.114}$$

Aplicando esta ecuación sobre (1.113) y, tras algunas manipulaciones algebraicas que omitiremos aquí, puede escribirse

$$\hat{\mathbf{s}}_{\text{MMSE}} = \left(\mathbf{H}^T\mathbf{V_R}^{-1}\mathbf{H} + \mathbf{V_S}^{-1}\right)^{-1}\mathbf{H}^T{\mathbf{V_R}}^{-1}\mathbf{x} \tag{1.115}$$

Observaciones independientes

Considérese el caso con $M = N$ (hay tantas observaciones como variables a estimar), $\mathbf{H} = \mathbf{I}$, siendo $\mathbf{I}$ la matriz unidad, y matrices de covarianzas diagonales $\mathbf{V_S} = \mathbf{D_S}$ y $\mathbf{V_R} = \mathbf{D_R}$ (lo que equivale a decir que todas las componentes de $\mathbf{S}$, y todas las de $\mathbf{R}$, son independientes). La particularización de (1.113) para este caso resulta en

$$\hat{\mathbf{s}}_{\text{MMSE}} = \mathbf{D_S}\left(\mathbf{D_S} + \mathbf{D_R}\right)^{-1}\mathbf{x} \tag{1.116}$$

La matriz $\mathbf{D_S}\left(\mathbf{D_S} + \mathbf{D_R}\right)^{-1}$ es una matriz diagonal, cuyo elemento i-ésimo de la diagonal es

$$\left[\mathbf{D_S}\left(\mathbf{D_S} + \mathbf{D_R}\right)^{-1}\right]_{ii} = \frac{v_{S_i}}{v_{R_i} + v_{S_i}}$$

donde v_{S_i} y v_{R_i} son las varianzas de e S_i y R_i respectivamente.

Por lo tanto, la estimación de la componente i-ésima del vector aleatorio $\mathbf{S}$ es

$$\hat{s}_{\text{MMSE},i} = \frac{v_{S_i}}{v_{R_i} + v_{S_i}} x_i \tag{1.117}$$

Nótese que este resultado implica que cada componente de $\mathbf{S}$ ha de ser estimada con un estimador similar al obtenido en el Ejemplo 1.17. Dicha conclusión era esperable, ya que el modelo de generación de observaciones en este caso puede escribirse como $\mathbf{X} = \mathbf{S} + \mathbf{R}$, siendo todas las componentes de $\mathbf{S}$ y $\mathbf{R}$ independientes entre sí. En otras palabras, el problema podría haber sido descompuesto en N problemas de estimación independientes equivalentes al estudiado en el Ejemplo 1.17.

Observaciones independientes de una misma variable aleatoria unidimensional

Consideremos la observación repetida de una variable aleatoria unidimensional S, estando sujeta cada medición a ruidos independientes de distinta varianza. Se pretende estimar el valor de S en base al conjunto de observaciones $\mathbf{X}$. Esto supone una particularización del modelo general estudiado en esta subsección, en el que

$$\mathbf{X} = \mathbf{1}\, S + \mathbf{R}$$

Es decir, $\mathbf{H} = \mathbf{1}$, siendo $\mathbf{1}$ un vector columna de dimensiones apropiadas con todas sus entradas iguales a 1, y siendo S una variable aleatoria unidimensional. El hecho de que las observaciones estén sujetas a ruidos indpendientes implica que la matriz de covarianza del ruido es diagonal, $\mathbf{V_R} = \mathbf{D_R}$ de componentes diagonales v_{R_i}.

Aplicando (1.115), se obtiene

$$\hat{\mathbf{s}}_{\text{MMSE}} = \frac{1}{v_S^{-1} + \mathbf{1}^T \mathbf{D}_\mathbf{R}^{-1} \mathbf{1}} \mathbf{1}^T \mathbf{D}_\mathbf{R}^{-1} \mathbf{x} \tag{1.118}$$

y, teniendo en cuenta que $\mathbf{D}_\mathbf{R}^{-1}$ es una matriz diagonal, se obtiene

$$\hat{s}_{\text{MMSE}} = \frac{1}{\sum_i v_{R,i}^{-1} + v_S^{-1}} \sum_i \frac{x_i}{v_{R,i}} \tag{1.119}$$

Respecto del resultado anterior, nótese que la estimación de S consiste en un promedio ponderado de las observaciones, asignando un mayor peso a aquellas observaciones contaminadas por una menor cantidad de ruido (i.e., con baja $v_{R,i}$).

1.8.2 Principio de Ortogonalidad. Interpretación geométrica

Una analogía que permite obtener algo más de intuición acerca del significado del Principio de Ortogonalidad obtenido en (1.87), así como del problema de estimación lineal de mínimo error cuadrático medio, consiste en asociar cada variable aleatoria unidimensional a un vector en un espacio euclídeo. La analogía, considerando el

caso en que todas las variables aleatorias tienen medias nulas, es como sigue (véase la Figura 1.9): cada variable aleatoria puede representarse como un vector en un espacio euclídeo, definiendo el producto escalar entre dos vectores en dicho espacio como su covarianza $\langle X_i, X_j \rangle = \mathbb{E}\{X_i X_j\}$ (recuérdese que estamos asumiendo medias nulas). De esta manera, la longitud del vector asociado a cada variable aleatoria es directamente la varianza de la variable, $\|X_i\| = \sqrt{\mathbb{E}\{X_i X_i\}}$. Puede comprobarse que, con estas definiciones, se satisfacen las necesarias correspondencias entre sumas y diferencias de variables aleatorias y sus correspondientes representaciones vectoriales.

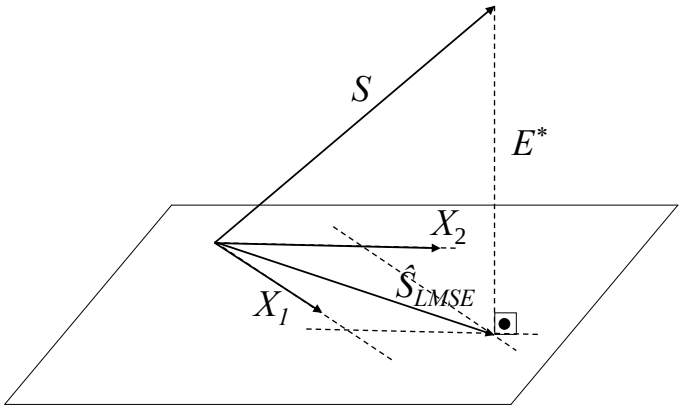


Fig. 1.9. Interpretación geométrica del *Principio de Ortogonalidad*.

Tanto las variables observables X_i como aquélla que deseamos estimar S se asocian por tanto a un vector en un espacio euclídeo. Ahora, si el objetivo es aproximar el valor de S como combinación lineal de las X_i, resulta claro que la estimación de S debe pertenecer al subespacio generado por las observaciones (un plano, para el caso de dos observaciones representado en la Figura 1.9). El objetivo de minimización de error cuadrático medio es análogo al de minimización de la norma del error ($\|E\|$), y sabemos que dicha norma se minimiza cuando el vector de error es ortogonal al subespacio generado por las X_i, y por tanto también ortogonal a todos los vectores de dicho espacio, incluidas cada una de las observaciones. Cuando recuperamos la interpretación en términos de variables aleatorias, dicha conclusión sigue siendo válida, sin más que argumentar en términos de ortogonalidad estadística en lugar de geométrica.

Un corolario interesante del Principio de Ortogonalidad, que también puede entenderse fácilmente a la vista de lo representado en la Figura 1.9, es que el error del

estimador lineal óptimo E^* también ha de ser ortogonal al propio estimador, $\hat{S}_{\text{LMSE}}$, por ser éste una combinación lineal de las observaciones y, por tanto, un vector en un subespacio ortogonal a E^*.

Para concluir la sección, conviene insistir en el hecho de que **todos estos resultados son válidos exclusivamente para el caso de estimación lineal de mínimo error cuadrático medio**.

1.9 Problemas

1.1. La distribución a posteriori de S dado X es

$$p_{S|X}(s|x) = x^2 \exp(-x^2 s), \qquad s \geq 0$$

Determine los estimadores $\hat{S}_{\text{MMSE}}$, $\hat{S}_{\text{MAD}}$ y $\hat{S}_{\text{MAP}}$.

1.2. Considere un problema de estimación caracterizado por la siguiente distribución a posteriori:

$$p_{S|X}(s|x) = x \exp(-xs), \quad s > 0 \tag{1.120}$$

Determine los estimadores $\hat{S}_{\text{MMSE}}$, $\hat{S}_{\text{MAD}}$ y $\hat{S}_{\text{MAP}}$.

1.3. Se desea estimar la v.a. S a partir de la observación de otra v.a. X mediante un estimador lineal de mínimo error cuadrático medio dado por la expresión:

$$\hat{S}_{\text{LMSE}} = w_0 + w_1 X$$

Sabiendo que $\mathbb{E}\{X\} = 1$, $\mathbb{E}\{S\} = 0$, $\mathbb{E}\{X^2\} = 2$, $\mathbb{E}\{S^2\} = 1$ y $\mathbb{E}\{SX\} = 1/2$, calcule:

a) Los valores de w_0 y w_1.

b) El error cuadrático medio del estimador, $\mathbb{E}\left\{\left(S - \hat{S}_{\text{LMSE}}\right)^2\right\}$.

1.4. Sean X y S dos variables aleatorias con d.d.p. conjunta

$$p_{X,S}(x,s) \begin{cases} 2 \; 0 < x < 1, 0 < s < x \\ 0 \text{ resto} \end{cases}$$

a) Calcule el estimador de mínimo error cuadrático medio de S dado X, $\hat{S}_{\text{MMSE}}$.
b) Calcule el sesgo del estimador $\hat{S}_{\text{MMSE}}$.

1.5. Se dispone de una imagen digitalizada de dimensiones 8×8 cuyos valores de luminancia son estadísticamente independientes y se distribuyen uniformemente entre 0 (blanco) y 1 (negro); se ha modificado dicha imagen aplicando sobre cada píxel una transformación de la forma $Y = X^r \; r > 0$, donde X es la v.a. asociada a los píxeles de la imágen original e Y la asociada a la imagen transformada. Obtenga la expresión que permite estimar por máxima verosimilitud el valor de r empleado en la transformación cuando se dispone de los 64 que componen la imagen transformada $\{y^{(k)}\}_{k=1}^{64}$, pero no se dispone de la imagen original.

1.6. Para el diseño de un sistema de comunicación se desea estimar la atenuación de señal entre el transmisor y el receptor, así como la potencia de ruido introducida por el canal cuando este ruido es gaussiano de media nula e independiente de la señal transmitida. Para ello, el transmisor envía una señal con una amplitud constante de 1 y el receptor recopila un conjunto de K observaciones disponibles a su entrada.

a) Estime por máxima verosimilitud la atenuación del canal, α, y la varianza del ruido, v_r, cuando las observaciones diponibles en el receptor son

$$\{0.55, 0.68, 0.27, 0.58, 0.53, 0.37, 0.45, 0.53, 0.86, 0.78\}.$$

b) Si el sistema se va a utilizar para la transimisión de señales digitales con una codificación unipolar (se emplea un nivel de señal A para transmitir el bit 1 y se mantiene el nivel de señal a 0 para la transmisión del bit 0), considerando equiprobabilididad entre símbolos, índique el mínimo nivel de señal que debe usarse en la codificación, $A_{\min}$, para garantizar un nivel de SNR en el receptor de 3 dB.

2

Aprendizaje Máquina

2.1 Principios generales del aprendizaje máquina

Como se ha estudiado en secciones anteriores, el diseño de estimadores y clasificadores que son capaces de aprender una función para la estimación o clasificación de cualquier nuevo punto $\mathbf{x}$ del espacio de observación (procedimiento denominado *inducción*) precisa de cierta información que relacione las observaciones y el valor a estimar (o la clase deseada). En los capítulos anteriores hemos asumido que dicha información estaba disponible gracias al conocimiento de determinadas distribuciones de probabilidad (enfoque analítico). Así, por ejemplo si en un determinado problema de estimación asumimos conocida $p_{S,\mathbf{X}}(s,\mathbf{x})$ disponemos de la caracterización más completa posible para el diseño óptimo de estimadores (de hecho, para el diseño de estimadores bayesianos resulta suficiente la distribución a posteriori de S).

En la práctica, existen un gran número de problemas en los que no se dispone del conocimiento estadístico necesario para llevar a cabo la tarea de estimación o clasificación de forma óptima. Sin embargo, si se dispone de *datos etiquetados*, $\{\mathbf{x}^{(k)}, s^{(k)}\}$, es decir, de un conjunto de observaciones para las cuales se conoce el valor de la variable objetivo, resulta posible utilizar dicha información para la construcción de estimadores o clasificadores siguiendo un enfoque conocido como *máquina* o de *aprendizaje automático*. Esto no es de extrañar, ya que puede entenderse que si el conjunto de datos $\{\mathbf{x}^{(k)}, s^{(k)}\}$ está compuesto por muestras i.i.d. de $p_{S,\mathbf{X}}(s,\mathbf{x})$, la información contenida en dicho conjunto de datos puede considerarse como una aproximación al propio conocimiento de la densidad de probabilidad, por lo que podrá utilizarse en la tarea de estimación. Obviamente, conforme el número de muestras disponibles crece, el conjunto de datos proporciona una información más completa acerca de la densidad de probabilidad conjunta real, por lo que el estimador o decisor construido se aproximará más al diseño óptimo analítico.

En esta sección presentamos algunos de los conceptos clave inherentes al diseño de estimadores y clasificadores a partir de datos. Cabe mencionar que existen al menos dos maneras de proceder a partir de dicho conjunto de datos:

- Los datos pueden utilizarse en primer lugar para obtener una aproximación de la densidad de probabilidad conjuntaNótese que en el caso de clasificación una alterna-

tiva frecuentemente utilizada consiste en la estimación de las verosimilitudes $p_{\mathbf{X}|H}(\mathbf{x}|h)$. Esto es mucho más complicado en el caso de estimación dado el carácter continuo de las variables objetivo, $\hat{p}_{S,\mathbf{X}}(s,\mathbf{x})$. Una vez se dispone de dicha estimación de la d.d.p., puede procederse siguiendo un enfoque analítico convencional. Esta aproximación se conoce habitualmente como semianalítica.

- Otra posibilidad es utilizar directamente los datos de entrenamiento para el proceso de estimación o clasificación, evitando la aproximación de densidad de probabiliidad alguna que es, en general un objetivo más complicado que la propia tarea de estimación o clasificación. Este enfoque es el que se suele asumir cuando se habla de Aprendizaje Máquina, y será el que estudiaremos de forma resumida en el presente capítulo.

Resulta pertinente plantearse la cuestión de cuál de los dos enfoques, analítico o máquina, resulta más potente para la resolución de problemas de aprendizaje. En principio, no hay situación más ventajosa que el conocimiento estadístico del problema. Sin embargo, en la práctica es habitual que dicha información no se conozca (o al menos no con exactitud), mientras que el acceso a un conjunto de datos etiquetado puede resultar más viable. Por ejemplo, si se considera un escenario de clasificación de imagen en diagnóstico médico, resulta evidente que la disponibilidad de un modelo estadístico preciso que relacione el valor de los píxeles de la imagen con la variable a estimar (e.g., nivel de respuesta a un determinado contraste) o la clase a predecir (e.g., presencia o no de tumores) es imposible, mientras que la construcción de un conjunto de pares etiquetados (conjunto de entrenamiento) únicamente requiere del etiquetado manual de imágenes concretas por parte de expertos, un procedimiento probablemente costoso, pero viable en cualquier caso.

En la literatura científica y técnica se viene realizando un gran esfuerzo en esta dirección a lo largo de las últimas décadas, disponiéndose actualmente de una amplia batería de métodos de aprendizaje automático. No es el objetivo de este capítulo cubrir siquiera un número reducido de las técnicas de aprendizaje propuestas, pero sí presentar de forma resumida algunos de los conceptos más importantes de dicho aprendizaje máquina, revisando únicamente algunas técnicas concretas a modo ilustrativa.

2.2 Métodos Paramétricos y no Paramétricos

- Paramétrico: Se propone un modelo en forma de una función parametrizada. Se trata de optimizar una determinada función de dichos datos que mida la discrepancia entre las variables objetivos disponibles y las que proporciona el modelo (función de coste basada en muestras, típicamente promedios muestrales). Adicionalmente, se pueden incluir términos de control de la generalización. En función del problema, podemos encontrar distintos métodos de optimización. Algunos de ellos proporcionan la solución óptima en modo bloque (i.e., existe una solución cerrada), mientras que otros proceden de manera iterativa. En la última parte del curso veremos algún ejemplo de dichos procesos iterativos en el caso particular de estimación.

- No Paramétrico: Son estrategias que no requieren la definición a priori de ningún tipo concreto de función que implemente el estimador o clasificador. Lo veremos con un ejemplo.

2.3 Estimación Máquina No Paramétrica: Método del vecino más próximo

$$\hat{s}(\mathbf{x}) = s^{(k^*)}$$

siendo

$$k^* = \arg\min_k \|\mathbf{x} - \mathbf{x}^{(k)}\|_2$$

2.4 Estimación Máquina Paramétrica: Regresión de Mínimos Cuadrados

$$\hat{s}(\mathbf{x}) = w_0 + \mathbf{w}^T\mathbf{x}$$

Los vectores óptimos se obtienen como

$$\begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} = (\mathbf{X}_{\mathrm{e}}^T\mathbf{X}_{\mathrm{e}})^{-1}\mathbf{X}_{\mathrm{e}}^T\mathbf{s}$$

con

$$\mathbf{X}_{\mathrm{e}} = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_N^{(1)} \\ 1 & x_1^{(2)} & \dots & x_N^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(K)} & \dots & x_N^{(K)} \end{bmatrix}$$

$$\mathbf{s} = [s^{(1)}, \cdots, s^{(K)}]^T$$

2.4.1 Modelos Semilineales

$$\hat{s} = w_0 + w_1 f_1(\mathbf{x}) + w_2 f_2(\mathbf{x}) + \cdots + w_{N'} f_{N'}(\mathbf{x}) = w_0 + w_1 y_1 + w_2 y_2 + \cdots + w_{N'} y_{N'}$$

$$\{\mathbf{x}^{(k)}, s^{(k)}\} \longrightarrow \{\mathbf{y}^{(k)}, s^{(k)}\}$$

$$\begin{bmatrix} w_0 \\ \mathbf{w} \end{bmatrix} = (\mathbf{Y}_{\mathrm{e}}^T\mathbf{Y}_{\mathrm{e}})^{-1}\mathbf{Y}_{\mathrm{e}}^T\mathbf{s}$$

con

$$\mathbf{Y}_{\mathrm{e}} = \begin{bmatrix} 1 & y_1^{(1)} & \dots & y_{N'}^{(1)} \\ 1 & y_1^{(2)} & \dots & y_{N'}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_1^{(K)} & \dots & y_{N'}^{(K)} \end{bmatrix}$$

2.5 Generalización

- Para el diseño bajo enfoque supervisado se dispone de un conjunto de entrenamiento con datos supervisados. No obstante, ha de tenerse presente que el objetivo es aplicar dicha máquina en nuevos datos, diferentes de los disponibles durante el entrenamiento.
- Generalización: La deseable propiedad de que la máquina proporcione una buena estimación/clasificación en datos diferentes de los del entrenamiento.
- Sobreajuste: El comportamiento indeseado que ocurre cuando la función de estimación o clasificación aprende las particularidades del conjunto de entrenamiento, debidas al ruido o al efecto del submuestreo, pero no extrapolables al problema real.
- Una forma de garantizar una adecuada generalización es mediante la aplicación de técnicas conocidas como de validación, de forma que un conjunto de datos se deja aparte para estimar el comportamiento de la máquina, y tener así una forma de predecir cuál va a ser el comportamiento en datos diferentes a los usados para el entrenamiento. La validación cruzada divide el conjunto de entrenamiento en varios subconjuntos, y promedia los resultados obtenidos al utilizar cada uno de ellos como de validación.
- En la práctica, el conjunto de test no es conocido, al menos no las etiquetas deseadas. No obstante, en los diseños de laboratorio es frecuente disponer de dichas etiquetas. Únicamente se pueden utilizar a los efectos de una evaluación final de los diferentes métodos y su comparación; en ningún caso deberían utilizarse durante el diseño.

3

Decisión analítica

3.1 Introducción al problema de decisión

En muchas situaciones debemos tomar decisiones, elegir entre varias alternativas diferentes, basándonos en observaciones o datos que tienen un comportamiento aleatorio. La teoría que permite encontrar soluciones a este tipo de problemas se la conoce como Teoría de la Decisión. Ejemplos de este tipo de situaciones nos las podemos encontrar:

- Al diseñar el receptor de un sistema de comunicaciones digitales, donde el transmisor envía un "0" o un "1" y, a partir de la señal presente a la entrada del receptor (observación), se debe decidir que símbolo ha sido transmitido.
- En un sistema radar donde a partir de la señal que llega al receptor radar se debe detectar si hay o no un blanco presente. Cuando la Teoría de la Decisión se aplica a este tipo de escenarios, suele denominarse Teoría de la Detección.
- En problemas de diagnóstico médico donde, por ejemplo, el médico examinando un electrocardiograma debe decidir si el paciente ha tenido o no un ataque al corazón.
- En problemas de clasificación de locutores donde a partir de la voz del locutor se debe decidir si el locutor es hombre o mujer, o bien decidir su nacionalidad (español, francés, alemán, ...) En el primer caso, donde sólo se decide entre dos posibles hipótesis, se dice que el problema de clasificación es binario, mientras que en el segundo caso (en el que hay más de dos hipótesis) se habla de problemas de clasificación multiclase.

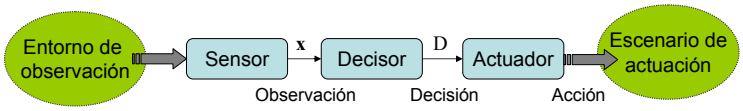


Fig. 3.1. Componentes de un sistema de decisión.

La figura 3.1 muestra los elementos principales que componen un sistema de decisión. Entre estos elementos nos encontramos:

- Las *hipótesis* que son aquellas posibles circunstancias, no observables, que dan lugar al problema de decisión. Dado que su valor es desconocido, se modelará como una variable aleatoria discreta. En general se considera que se dispone de un conjunto de L hipótesis exhaustivas y mutuamente exclusivas (es decir, que una y solamente una de las hipótesis es cierta), y se representan mediante la variable aleatoria H definida sobre $\{0, 1, \ldots, L-1\}$ de tal modo que $H = h$ si y sólo si la hipótesis h-ésima es correcta. Así, en el ejemplo del problema de comunicaciones tendríamos dos posibles hipótesis $H = 0$ (se transmitió el símbolo "0") o $H = 1$ (se transmitió el símbolo "1"); mientras que en el problema de detección de la nacionalidad del locutor habría tantas hipótesis como posibles nacionalidades a clasificar.
- Las *observaciones* que estarán contenidas en el vector aleatorio $\mathbf{X}$. Este vector estará formado por aquella información capturada por el *sensor* del *entorno* y que puede resultar útil para la decisión (debe haber una relación estadística entre las hipótesis y el vector de observaciones para que el problema de decisión tenga sentido). En un caso general la observación es un vector de números reales, es decir, $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^N$. Así, por ejemplo, en el sistema de comunicaciones las observaciones son las señales captadas en el receptor y cuyo valor dependerá, en un caso general, del símbolo transmitido y del ruido aleatorio introducido por el canal.
- El *decisor* que será el elemento del sistema que toma la *decisión* $D = \phi(\mathbf{X})$ tras observar $\mathbf{X}$, siendo ϕ la *función de decisión*. La decisión es una variable aleatoria discreta perteneciente a un alfabeto finito que, por simplicidad, codificaremos mediante números enteros consecutivos, $\mathcal{M} = \{0, 1, \ldots, M-1\}$. Cada una de las decisiones posibles (es decir, cada uno de los elementos de $\mathcal{M}$) se denominan *categorías* o *clases*. Por último, es importante destacar que la función de decisión tiene un carácter determinista, es decir, para un valor dado de $\mathbf{x}$ el decisor siempre decidirá la misma categoría, $d = \phi(\mathbf{x})$. Además, por cada posible valor de $\mathbf{x}$ está función sólo le asigna una categoría, dando así lugar a las llamadas *regiones de decisión* en las que nos centraremos más adelante.
- El último de los elementos del sistema es el *actuador*, encargado de ejecutar sobre el escenario de actuación las acciones derivadas de la decisión D. El decisor debe diseñarse teniendo en cuenta las consecuencias derivadas de estas acciones. Así, por ejemplo, en el caso del sistema de comunicación el receptor suele diseñarse intentando minimizar la probabilidad de error; por el contrario, en el caso de diagnóstico médico habría que considerar los efectos de dar un diagnóstico erróneo, teniendo en cuenta que las consecuencias derivadas de no detectar un ataque de corazón, si este se ha producido, son muy diferentes del caso opuesto, haberlo diagnosticado cuando no ha ocurrido.

El esquema anterior es muy general, y representa de forma simplificada muchos procesos de toma decisiones en hombres y máquinas. Aunque el sensor, el decisor y el actuador pueden ser manuales, o automáticos, el objetivo principal de este tema

es determinar procedimientos automáticos de decisión (independientemente de la naturaleza del sensor o el actuador).

3.1.1 Regiones de decisión

Como se acaba de indicar, la función de decisión, ϕ, al asignar una y sólo una categoría a cada posible observación, $\mathbf{x}$, divide el espacio muestral en regiones, una por cada categoría. Llamaremos *región de decisión* de la categoría d al conjunto definido por

$$\mathcal{X}_d = \{\mathbf{x} \in \mathcal{X} | \phi(\mathbf{x}) = d\} \tag{3.1}$$

y llamaremos *fronteras de decisión* a aquellas que limitan las regiones de decisión.

Por tanto, todo decisor induce la partición $\mathcal{X} = \bigcup_{d=0}^{M-1} \mathcal{X}_d$. Obsérvese también que las regiones de decisión caracterizan completamente un decisor. Podemos diseñar un decisor, bien determinando el valor de ϕ para todo $\mathbf{x} \in \mathcal{X}$, o bien especificando las regiones $\mathcal{X}_d$ de una partición sobre $\mathcal{X}$.

Example 3.1. El decisor $\phi(x) = u(x^2 - 1)$ (donde $u(.)$ es la función escalón), definido sobre $\mathcal{X} = \mathbb{R}$, se caracteriza por las regiones de decisión:

$$\mathcal{X}_0 = \{x \in \mathbb{R} | x^2 - 1 < 0\} = (-1, 1) \tag{3.2}$$

$$\mathcal{X}_1 = \{x \in \mathbb{R} | x^2 - 1 \geq 0\} = (-\infty, -1] \cup [1, \infty) \tag{3.3}$$

(donde hemos supuesto $u(0) = 1$).

Example 3.2. El decisor $\phi(\mathbf{x}) = \underset{i}{\operatorname{argmin}}\, y_i(\mathbf{x})$, definido sobre $\mathcal{X} = [0, 1]^2$, siendo

$$y_0 = \|\mathbf{x}\|^2 \tag{3.4}$$

$$y_1 = x_1 - x_0 + 1 \tag{3.5}$$

$$y_2 = x_0 - x_1 + 1 \tag{3.6}$$

se caracteriza por las regiones de decisión que se muestran en la fig. 3.2.

3.1.2 Diseño de decisores

El diseño de un decisor depende del tipo de información disponible acerca del problema de decisión. Discutiremos dos familias principales de procedimientos de diseño:

- Métodos **analíticos**: basados en el uso de información estadística del problema. Pueden emplearse cuando, por la naturaleza del problema, es posible determinar un modelo probabilístico de las variables relevantes del problema (entre ellas, las observaciones)
- Métodos **máquina**: basados en el uso de información empírica sobre el problema. Pueden emplearse cuando no se dispone de un modelo probabilístico fiable pero, a cambio, se tienen datos históricos útiles que proporcionan pistas para diseñar el sistema.

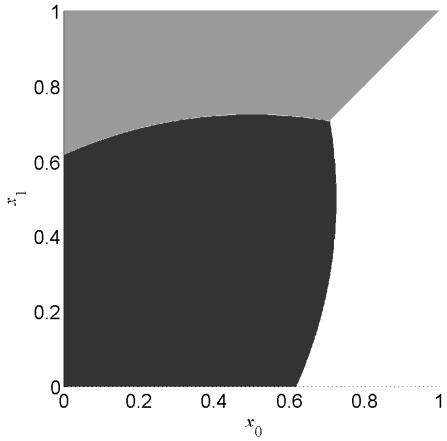


Fig. 3.2. Regiones de decisión del decisor del ejemplo 3.2: $\mathcal{X}_0$ (negro), $\mathcal{X}_1$ (gris) y $\mathcal{X}_2$ (blanco)

3.2 Diseño analítico de decisores

3.2.1 Modelado estadístico de los problemas de decisión

Antes de entrar a describir la teoría de la decisión es conveniente presentar las funciones de probabilidad que caracterizan estadísticamente la relación existente entre observaciones y las hipótesis:

- En primer lugar, la **verosimilitud** de H viene dada por $p_{\mathbf{X}|H}(\mathbf{x}|h)$, y caracteriza probabilísticamente la generación de las observaciones para cada posible hipótesis. Para un problema de clasificación con L hipótesis, nos encontraremos con L funciones de densidad de probabilidad que definen la verosimilitud sobre cada posible hipótesis: $p_{\mathbf{X}|H}(\mathbf{x}|0), p_{\mathbf{X}|H}(\mathbf{x}|1), \ldots p_{\mathbf{X}|H}(\mathbf{x}|L-1)$.
- La **distribución marginal o a priori** de H, denotada como $P_H(h)$. Nótese que como H es una v.a. discreta debe verificarse que $\sum_{h=0}^{L-1} P_H(h) = 1$.
- La **distribución de las observaciones**, $p_{\mathbf{X}}(\mathbf{x})$, que modela la densidad de probabilidad de $\mathbf{X}$ en el punto $\mathbf{x}$.
- La **distribución conjunta** de $\mathbf{X}$ y H: $p_{\mathbf{X},H}(\mathbf{x}, h) = p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)$
- La **distribución a posteriori** de H: $P_{H|\mathbf{X}}(h|\mathbf{x})$ que indica con que probabilidad ocurre cada hipótesis para un valor dado de $\mathbf{x}$.

Es importante resaltar que la información disponible para el diseño del estimador puede ser diferente en cada situación concreta. Una situación habitual, por estar relacionada con el propio proceso físico de generación de las observaciones, es aquélla en la que se dispone de la verosimilitud y la probabilidad a priori de cada hipótesis, $P_H(h)$. Nótese que a partir de ellas puede obtenerse cualquier otra distribución, así, por ejemplo, mediante el Teorema de Bayes se puede obtener la distribución a posteriori de las hipótesis como:

$$P_{H|\mathbf{X}}(h|\mathbf{x}) = \frac{p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)}{\sum_{h=0}^{L-1} p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)} \tag{3.7}$$

3.2.2 Riesgo

¿Cómo determinar si un decisor es mejor que otro? Para ello necesitamos una medida que permita evaluar las prestaciones de cualquier decisor. La idea germinal de la teoría de la decisión es la siguiente: supongamos que es posible cuantificar las consecuencias derivadas de cada decisión (o, más precisamente, de la acción consecuente a cada decisión) mediante una función de coste $c(D,H) \in \mathbb{R}$ que asigna una penalización c_{dh}, con $c_{dh} > c_{hh} \geq 0,\ \forall d \neq h$, al hecho de decidir $D = d$ cuando la hipótesis es $H = h$.

Una vez definida la política de costes, pueden medirse las prestaciones de un decisor dado por una función de decisión ϕ mediante su coste medio, también denominado como **riesgo**

$$\begin{aligned} r_\phi =& \mathbb{E}\{c(D,H)\} = \sum_{d=0}^{M-1}\sum_{h=0}^{L-1} c_{dh} P\{D=d, H=h\} \\ =& \sum_{d=0}^{M-1}\sum_{h=0}^{L-1} c_{dh} P_H(h) P_{D|H}(d|h) \end{aligned} \tag{3.8}$$

Teniendo en cuenta que $\phi(\mathbf{x}) = d$ cuando la observación pertenece a la región de decisión de $D = d$, es decir, $\mathbf{x} \in \mathcal{X}_d$, el término $P_{D|H}(d|h)$ se puede calcular como

$$P_{D|H}(d|h) = P\{X \in \mathcal{X}_d | H = h\} = \int_{\mathcal{X}_d} p_{\mathbf{X}|H}(\mathbf{x}|h) d\mathbf{x} \tag{3.9}$$

Example 3.3. Se tiene un problema de decisión multiclase con tres hipótesis cuyas verosimilitudes son:

$$\begin{array}{ll} p_{X|H}(x|0) = 1 & 0 < x < 1 \\ p_{X|H}(x|1) = 2(1-x) & 0 < x < 1 \\ p_{X|H}(x|2) = 2x & 0 < x < 1 \end{array}$$

sabiendo que las probabilidades a priori de las hipótesis son: $P_H(0) = 0.4$ y $P_H(1) = P_H(2) = 0.3$ y la política de costes viene dada por $c_{hh} = 0,\ h = 0,1,2$ y $c_{hd} = 1,\ h \neq d$, obtenga el riesgo del decisor:

$$\phi(x) = \begin{cases} 1, & x < 0.5 \\ 2, & x > 0.5 \end{cases}$$

Aplicando la expresión (3.8) a este problema se tiene:

$$\begin{aligned} r_\phi = & c_{10}P_H(0)P_{D|H}(1|0) + c_{20}P_H(0)P_{D|H}(2|0) + c_{01}P_H(1)P_{D|H}(0|1) \\ & + c_{21}P_H(1)P_{D|H}(2|1) + c_{02}P_H(2)P_{D|H}(0|2) + c_{12}P_H(2)P_{D|H}(1|2) \end{aligned}$$

donde se pueden calcular los términos $P_{D|H}(d|h)$ aplicando (3.9)

$$\begin{aligned}
&P_{D|H}(0|1) = P_{D|H}(0|2) = 0\\
&P_{D|H}(1|0) = \textstyle\int_{\mathcal{X}_1} p_{X|H}(x|0)\, dx = \int_0^{0.5} 1\, dx = 0.5\\
&P_{D|H}(2|0) = \textstyle\int_{\mathcal{X}_2} p_{X|H}(x|0)\, dx = \int_{0.5}^1 1\, dx = 0.5\\
&P_{D|H}(1|2) = \textstyle\int_{\mathcal{X}_1} p_{X|H}(x|2)\, dx = \int_0^{0.5} 2x\, dx = 0.25\\
&P_{D|H}(2|1) = \textstyle\int_{\mathcal{X}_2} p_{X|H}(x|1)\, dx = \int_{0.5}^1 2(1-x)\, dx = 0.25
\end{aligned}$$

y sustituyendo se llega a

$$r_\phi = 0.4 \cdot 0.5 + 0.4 \cdot 0.5 + 0.3 \cdot 0.25 + 0.3 \cdot 0.25 = 0.55$$

De manera alternativa, se puede evaluar la calidad de la decisión d para una observación $\mathbf{x}$ mediante el coste medio dado la observación o el **riesgo condicional**

$$\mathbb{E}\{c(d,H)|\mathbf{x}\} = \sum_{h=0}^{L-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}) \tag{3.10}$$

el cual puede relacionarse con el riesgo o coste medio del decisor a través de la expresión:

$$r_\phi = \mathbb{E}\{c(D,H)\} = \int \mathbb{E}\{c(d,H)|\mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{3.11}$$

Esta relación entre ambos riesgos pone de manifiesto que ambas funciones de coste son equivalentes de cara a diseñar el decisor que proporcione el mínimo riesgo, ya que el riesgo condicional estaría minimizando para cada valor de $\mathbf{x}$ el coste medio del decisor.

Example 3.4. Continuando con el ejemplo 3.3 se podría calcular el riesgo condicional de cada decisión como

$$\mathbb{E}\{c(d,H)|x\} = c_{d0} P_{H|X}(0|x) + c_{d1} P_{H|X}(1|x) + c_{d2} P_{H|X}(2|x)$$

donde se pueden obtener las distribuciones a posteriori mediante la aplicación del Teorema de Bayes

$$\begin{aligned}
P_{H|X}(0|x) &= \frac{p_{X|H}(x|0)P_H(0)}{\sum_{h=0}^2 p_{X|H}(x|h)P_H(h)} = \frac{1 \cdot 0.4}{1 \cdot 0.4 + 2(1-x) \cdot 0.3 + 2x \cdot 0.3} = 0.4\\
P_{H|X}(1|x) &= \frac{p_{X|H}(x|1)P_H(1)}{\sum_{h=0}^2 p_{X|H}(x|h)P_H(h)} = \frac{2(1-x) \cdot 0.2}{1} = 0.6(1-x)\\
P_{H|X}(2|x) &= \frac{p_{X|H}(x|2)P_H(2)}{\sum_{h=0}^2 p_{X|H}(x|h)P_H(h)} = \frac{2x \cdot 0.3}{1} = 0.6x
\end{aligned}$$

y se llega a:

- si $d = 0$:

$$\begin{aligned}\mathbb{E}\{c(0,H)|x\} &= c_{00}P_{H|X}(0|x) + c_{01}P_{H|X}(1|x) + c_{02}P_{H|X}(2|x)\\ &= 0 \cdot 0.4 + 1 \cdot 0.6(1-x) + 1 \cdot 0.6x = 0.6\end{aligned}$$

- si $d = 1$:

$$\begin{aligned}\mathbb{E}\{c(1,H)|x\} &= c_{10}P_{H|X}(0|x) + c_{11}P_{H|X}(1|x) + c_{12}P_{H|X}(2|x)\\ &= 1 \cdot 0.4 + 0 \cdot 0.6(1-x) + 1 \cdot 0.6x = 0.4 + 0.6x\end{aligned}$$

- si $d = 2$:

$$\begin{aligned}\mathbb{E}\{c(2,H)|x\} &= c_{20}P_{H|X}(0|x) + c_{21}P_{H|X}(1|x) + c_{22}P_{H|X}(2|x)\\ &= 1 \cdot 0.4 + 1 \cdot 0.6(1-x) + 0 \cdot 0.6x = 1 - 0.6x\end{aligned}$$

3.2.3 Teoría bayesiana de la decisión

La ecuación (3.10) nos proporciona de forma inmediata una regla para diseñar el decisor óptimo: la decisión $D = d$ será la óptima si es la que proporciona el menor riesgo condicional

$$\phi^*(\mathbf{x}) = \underset{d}{\operatorname{argmin}} \sum_{h=0}^{L-1} c_{dh} P_{H|\mathbf{X}}(h|\mathbf{x}) \tag{3.12}$$

A modo de ejemplo, la Figura 3.3 muestra la estructura del decisor de mínimo riesgo para un problema con dos hipótesis y tres alternativas.

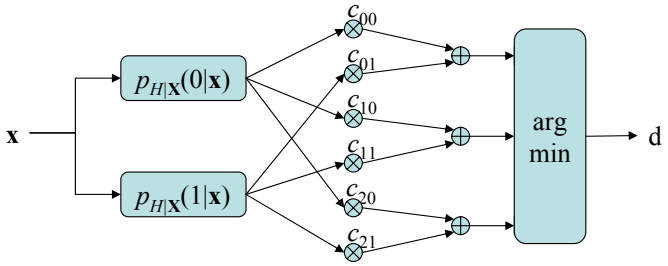


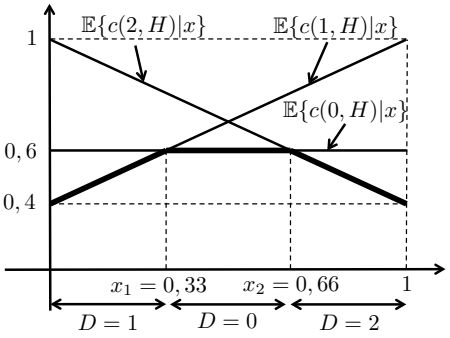
Fig. 3.3. Estructura del decisor de mínimo riesgo con dos hipótesis y tres alternativas.

Example 3.5. Obtenga el decisor de mínimo riesgo del problema de decisión dado en el ejemplo 3.3.

Partiendo de las expresiones del riesgo condicional obtenidas en el ejemplo 3.4,

$$\mathbb{E}\{c(0,H)|x\} = 0.6$$
$$\mathbb{E}\{c(1,H)|x\} = 0.4 + 0.6x$$
$$\mathbb{E}\{c(2,H)|x\} = 1 - 0.6x$$

sólo hay que analizar para cada valor de x que término es menor:



Calculo de x_1:
$0,4 + 0,6x_1 = 0,6$
$x_1 = 0,33$

Calculo de x_2:
$1 - 0,6x_2 = 0,6$
$x_2 = 0,66$

se tiene que el decisor bayesiano es:

$$\phi(x) = \begin{cases} 1, & 0 < x < 0.33 \\ 0, & 0.33 < x < 0.66 \\ 2, & 0.66 < x < 1 \end{cases}$$

Podemos obtener una expresión alternativa para el decisor óptimo aplicando la Regla de Bayes sobre las probablidades a posteriori,

$$P_{H|\mathbf{X}}(h|\mathbf{x}) = \frac{p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)}{p_{\mathbf{X}}(\mathbf{x})} \tag{3.13}$$

sustituyendo (3.13) en (3.15) se llega a

$$\phi^*(\mathbf{x}) = \underset{d}{\operatorname{argmin}} \sum_{h=1}^{L} c_{dh} \frac{p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h)}{p_{\mathbf{X}}(\mathbf{x})} \tag{3.14}$$

y teniendo en cuenta que el denominador no afecta a la decisión, (3.14) puede reducirse a

$$\phi^*(\mathbf{x}) = \underset{d}{\operatorname{argmin}} \sum_{h=0}^{L-1} c_{dh} p_{\mathbf{X}|H}(\mathbf{x}|h)P_H(h) \tag{3.15}$$

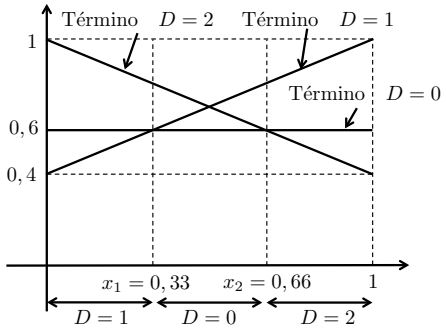
El uso de un criterio u otro (el dado por (3.12) o (3.15)) simplemente dependerá de la información de la que se tenga de partida, ya que en el primer caso se necesita disponer de las distribuciones de verosimilitud de H y las probabilidades a priori, mientras que en el segundo caso es necesario conocer la distribución a posteriori. Lógicamente, a partir de la función de verosimilitud y las probabilidades a priori se podría obtener la distribución a posteriori y aplicar la expresión (3.12) para el diseño del decisor.

Example 3.6. Aplicando la expresión (3.15), obtenga el decisor bayesiano correspondiente al problema de clasificación dado en el ejemplo 3.3

Para aplicar el criterio dado en (3.15) se debe calcular, en primer lugar, el término $\sum_{h=0}^{L-1} c_{dh} p_{\mathbf{X}|H}(\mathbf{x}|h) P_H(h)$ sobre cada posible decisión

- Si $D = 0$:
  $\sum_{h=0}^{2} c_{0h} p_{X|H}(x|h) P_H(h) = 0 \cdot 1 \cdot 0.4 + 1 \cdot 2(1-x) \cdot 0.3 + 1 \cdot 2x \cdot 0.3 = 0.6.$
- Si $D = 1$:
  $\sum_{h=0}^{2} c_{1h} p_{X|H}(x|h) P_H(h) = 1 \cdot 1 \cdot 0.4 + 0 \cdot 2(1-x) \cdot 0.3 + 1 \cdot 2x \cdot 0.3 = 0.4 + 0.6x.$
- Si $D = 2$:
  $\sum_{h=0}^{2} c_{2h} p_{X|H}(x|h) P_H(h) = 1 \cdot 1 \cdot 0.4 + 1 \cdot 2(1-x) \cdot 0.3 + 0 \cdot 2x \cdot 0.3 = 1 - 0.6x.$

Y analizando para cada valor de x que decisión incurre en un menor coste, tal y como puede hacerse representado cada término en función de x



Calculo de x_1:
$0,4 + 0,6x_1 = 0,6$
$x_1 = 0,33$

Calculo de x_2:
$1 - 0,6x_2 = 0,6$
$x_2 = 0,66$

se obtiene que el decisor bayesiano viene dado por:

$$\phi(x) = \begin{cases} 1, & 0 < x < 0.33 \\ 0, & 0.33 < x < 0.66 \\ 2, & 0.66 < x < 1 \end{cases}$$

que, como era de esperar, coincide con la expresión del decisor calculado en 3.5.

Decisión MAP

Posiblemente el caso más frecuente en problemas de decisión basadas en hipótesis es aquel en el que el número de alternativas es igual al de hipótesis, y el actuador se diseña para ejecutar las acciones más adecuadas para cada hipótesis. El decisor, por tanto, *debería* decidir $D = h$ cuando la hipótesis correcta es $H = h$. En tal caso, se calificará la decisión como un acierto, y se hablará de fallo o error cuando $D \neq h$. Dado que, en el momento de la decisión, la hipótesis correcta es desconocida, en general no es posible acertar sistemáticamente, y se producirán errores.

Si se supone que todos los tipos de errores son igualmente indeseados, mientras que los aciertos no penalizan, se puede escribir

$$c_{dh} = \begin{bmatrix} 1, \text{ si } d \neq h \\ 0, \text{ si } d = h \end{bmatrix} = 1 - \delta_{d-h} \tag{3.16}$$

en cuyo caso, el coste medio en (3.8) se reduce a

$$r_\phi = \sum_{d \neq h} P\{D = d, H = h\} = P\{D \neq H\} \tag{3.17}$$

que es la **probabilidad de error**. Por tanto, el decisor óptimo para costes (3.16) es el de mínima probabilidad de error. Para obtenerlo, podemos calcular en primer lugar el riesgo condicional. Usando (3.10), resulta

$$\mathbb{E}\{C(d,H)|\mathbf{x}\} = \sum_{h \neq d} P_{H|\mathbf{X}}(h|\mathbf{x}) = 1 - P_{H|\mathbf{X}}(d|\mathbf{x}) \tag{3.18}$$

luego, en virtud de (3.12), se obtiene el decisor

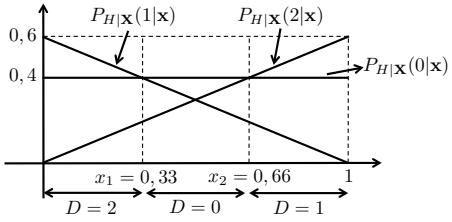
$$\phi_{\text{MAP}}(\mathbf{x}) = \underset{h}{\operatorname{argmax}}\, P_{H|\mathbf{X}}(h|\mathbf{x}) \tag{3.19}$$

En consecuencia, el decisor de mínima probabilidad de error es aquel que toma, para cada observación $\mathbf{x}$ la decisión asociada a la hipótesis más probable *a posteriori* (es decir, dado $\mathbf{x}$). Por este motivo, se le conoce como decisor MAP o de *Maximo A Posteriori*.

Example 3.7. Obtenga el decisor MAP del problema de decisión del ejemplo 3.3.

Para este problema se obtuvieron las siguiente distribuciones a posteriori:

$$P_{H|X}(0|x) = 0.4$$
$$P_{H|X}(1|x) = 0.6(1-x)$$
$$P_{H|X}(2|x) = 0.6x$$



Lo que proporciona el decisor:

$$\phi_{\text{MAP}}(x) = \begin{cases} 1, & 0 < x < 0.33 \\ 0, & 0.33 < x < 0.66 \\ 2, & 0.66 < x < 1 \end{cases}$$

que como puede verse coincide con el decisor bayesiano, ya que en este ejemplo la política de costes no penaliza a los aciertos y los errores penalizan por igual.

3.2.4 Decisión ML

Se denomina decisor de máxima verosimilitud, o ML (*Maximum Likelihood*) a aquél que se rige por la regla de decisión

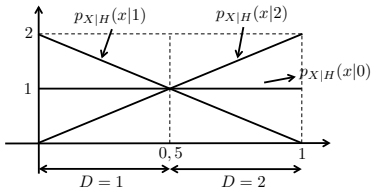
$$\phi^*_{\text{ML}}(\mathbf{x}) = \underset{h}{\operatorname{argmax}}\, p_{\mathbf{X}|H}(\mathbf{x}|h) \tag{3.20}$$

Aplicando la regla de Bayes sobre las probabilidades a posteriori en (3.19) se puede comprobar fácilmente que el decisor ML y el decisor MAP son equivalentes cuando todas las hipótesis son equiprobables (*a priori*), es decir $P_H(h) = 1/L$. Pero, de forma general, el decisor ML y el decisor MAP son distintos y ofrecen prestaciones diferenciadas. En particular, el decisor ML es una buena alternativa cuando no se conocen las probabilidades a priori (ni tampoco a posteriori) de las hipótesis.

Example 3.8. Continuando con el problema dado en el ejemplo 3.3, obtenga el decisor ML.

Para obtener el decisor simplemente hay que analizar para cada valor de x que verosimilud toma una mayor probabilidad:

$$\begin{aligned} p_{X|H}(x|0) &= 1 && 0 < x < 1 \\ p_{X|H}(x|1) &= 2(1-x) && 0 < x < 1 \\ p_{X|H}(x|2) &= 2x && 0 < x < 1 \end{aligned}$$



Su representación, directamente muestra las regiones de decisión:

$$\phi(x) = \begin{cases} 1, & 0 < x < 0.5 \\ 2, & 0.5 < x < 1 \end{cases}$$

Y nunca se decidiría $D = 0$. Nótese que en este caso el decisor obtenido difiere de la solución proporcionada por el decisor bayesiano, lo que se debe a la no equiprobabilidad entre las hipótesis.

3.3 Decisores binarios

En esta sección se analizaremos con detalle el diseño de decisores binarios ($M = 2$) basados en hipótesis binarias ($L = 2$). Interpretaremos los eventos $\{D = 0, H = 0\}$ y $\{D = 1, H = 1\}$ como aciertos y los eventos $\{D = 0, H = 1\}$ y $\{D = 1, H = 0\}$ como errores. Ademas, utilizaremos la siguiente terminología[1]

- Evento **Detección**: $\{D = 1, H = 1\}$
- Evento **Pérdida**: $\{D = 0, H = 1\}$
- Evento **Falsa Alarma**: $\{D = 1, H = 0\}$

Estudiaremos el decisor de mínimo riesgo, que no es más que una particularización del decisor bayesiano obtenido en la sección anterior, pero veremos también que existen otras alternativas al diseño de decisores que no están basadas en la medida de riesgo.

[1] Esta terminología proceden originalmente de aplicaciones radar, en las que la hipótesis $H = 1$ denota la presencia de un blanco que se pretende detectar, pero se ha extendido a otros ámbitos, incluso aunque el significado original puede carecer de sentido.

3.3.1 Riesgo de un decisor binario

En el caso binario nos encontraremos que la política de costes viene definida por cuatro términos: c_{00}, c_{11}, c_{10}, c_{01}, en los que, para cada hipótesis, el coste de un error es mayor que el de un acierto, es decir, $c_{10} > c_{00}$ y $c_{01} > c_{11}$. El riesgo asociado a un decisor ϕ binario, dado por (3.8) puede escribirse en tal caso como

$$\begin{aligned} r_\phi =& c_{00}P\{D=0,H=0\} + c_{01}P\{D=0,H=1\} \\ &+ c_{10}P\{D=1,H=0\} + c_{11}P\{D=1,H=1\} \\ =& c_{00}P_H(0)P_{D|H}(0|0) + c_{01}P_H(1)P_{D|H}(0|1) \\ &+ c_{10}P_H(0)P_{D|H}(1|0) + c_{11}P_H(1)P_{D|H}(1|1) \end{aligned} \tag{3.21}$$

Llamaremos **Probabilidad de Falsa Alarma**[2] a

$$P_{\text{FA}} = P_{D|H}(1|0) = \int_{\mathcal{X}_1} p_{\mathbf{X}|H}(\mathbf{x}|0)d\mathbf{x} \tag{3.22}$$

y **Probabilidad de Pérdida** a

$$P_{\text{M}} = P_{D|H}(0|1) = \int_{\mathcal{X}_0} p_{\mathbf{X}|H}(\mathbf{x}|1)d\mathbf{x} \tag{3.23}$$

De acuerdo con esto, podemos escribir

$$\begin{aligned} r_\phi =& c_{00}P_H(0)(1-P_{\text{FA}}) + c_{01}P_H(1) + c_{10}P_H(0)P_{\text{FA}} + c_{11}P_H(1)(1-P_{\text{M}}) \\ =& (c_{01}-c_{11})P_H(1)P_{\text{M}} + (c_{10}-c_{00})P_H(0)P_{\text{FA}} + (c_{00}P_H(0) + c_{11}P_H(1)) \end{aligned} \tag{3.24}$$

La expresión anterior muestra que el riesgo de un decisor es suma de tres componentes:

- $(c_{00}P_H(0)+c_{11}P_H(1))$ es el riesgo mínimo del decisor ideal, aquél con $P_{\text{M}} = 0$ y $P_{\text{FA}} = 0$ que acierta con probabilidad 1.
- $(c_{01} - c_{11})P_H(1)P_{\text{M}}$ es el incremento del riesgo que producen los errores de pérdida.
- $(c_{10}-c_{00})P_H(0)P_{\text{FA}}$ es el incremento del riesgo que producen las falsas alarmas.

Obsérvese que el decisor ideal es, en general, irrealizable, porque si las verosimilitudes de las hipótesis están solapadas, no es posible evitar que se produzcan errores. El decisor óptimo será aquel que encuentre un buen compromiso entre errores de pérdida y falsas alarmas, de tal manera que el riesgo en (3.21) sea mínimo.

[2] Observe que se utiliza la denominación "probabilidad de falsa alarma" para la probabilidad *condicional* del evento $D = 1$ dado $H = 0$, y por tanto no debe confundirse con la probabilidad del evento falsa alarma, $P\{D = 1, H = 0\}$. Comentario análogo aplica a la probabilidad de pérdida.

3.3.2 Función discriminante

De forma general, cualquier decisión binaria se puede expresar como el resultado de comparar cierta función de la observación con un umbral, es decir

$$g(\mathbf{x}) \underset{D=0}{\overset{D=1}{\gtrless}} \eta \tag{3.25}$$

La función g se conoce como **función discriminante**.

Dado que la observación es una variable aleatoria, la función discriminante define una nueva variable aleatoria

$$\Lambda = g(\mathbf{X}) \tag{3.26}$$

que puede resultar útil para el cálculo de las probabilidades de error, falsa alarma y pérdida. Así, por ejemplo, la probabilidad de falsa alarma será

$$P_{\text{FA}} = P_{D|H}(1|0) = P\{\Lambda > \eta | H = 0\} = \int_{\eta}^{\infty} p_{\Lambda|H}(\lambda|0) d\lambda \tag{3.27}$$

Analogamente,

$$P_{\text{M}} = P_{D|H}(0|1) = P\{\Lambda > \eta | H = 1\} = \int_{-\infty}^{\eta} p_{\Lambda|H}(\lambda|1) d\lambda \tag{3.28}$$

Cuando la observación es un vector, las integrales escalares en (3.27) y (3.28) son una buena alternativa a las integrales multidimensionales en (3.22) y (3.23), aunque requieren determinar previamente $p_{\Lambda|H}$.

Example 3.9. Consideremos el problema de decisión $\mathbf{X} = (X_1, X_2)$, dado por las verosimilitudes gaussianas

$$p_{\mathbf{X}|H}(\mathbf{x}|h) = \frac{1}{2\pi} \exp\left(-\frac{1}{2}((x_1 - h)^2 + (x_2 - h)^2)\right) \tag{3.29}$$

con $h = 0, 1$. Determinaremos la probabilidad de falsa alarma del decisor dado por la función discriminante $g(\mathbf{x}) = x_1 + x_2$ y umbral $\eta = 1$, es decir,

$$x_1 + x_2 \underset{D=0}{\overset{D=1}{\gtrless}} 1 \tag{3.30}$$

Dado que $\mathcal{X}_1 = \{(x_1, x_2) | x_1 + x_2 > 1\}$, la probabilidad de falsa alarma de este decisor puede obtenerse como

$$P_{\text{FA}} = \int_{-\infty}^{\infty} \int_{1-x_2}^{\infty} p_{\mathbf{X}|H}(\mathbf{x}|0) dx_1 dx_2 \tag{3.31}$$

que implica calcular una integral bidimensional. Sin embargo, dado que $\Lambda = g(\mathbf{X}) = X_1 + X_2$ es suma de variables gaussianas (bajo cualquiera de las hipótesis), $p_{\Lambda|H}(\lambda|h)$ también es gaussiana. En, particular, $p_{\Lambda|H}(\lambda|0)$ es una gaussiana de media

$$\mathbb{E}\{\Lambda|H=0\} = \mathbb{E}\{X_1|H=0\} + \mathbb{E}\{X_2|H=0\} = 0 \tag{3.32}$$

y varianza

$$\mathbb{E}\{\Lambda^2|H=0\} = \mathbb{E}\{(X_1+X_2)^2|H=0\} = 2 \tag{3.33}$$

Por tanto,

$$P_{\text{FA}} = \int_1^\infty p_{\mathbf{\Lambda}|H}(\lambda|0)d\lambda = \frac{1}{\sqrt{4\pi}}\int_1^\infty \exp\left(-\frac{1}{4}\lambda^2\right)d\lambda \approx 0.2398 \tag{3.34}$$

3.3.3 Decisores binarios de mínimo riesgo

El diseño de un decisor binario puede plantearse como el de determinar una función discriminante y un umbral de tal modo que la regla (3.25) tenga riesgo mínimo. Para ello, particularizaremos el resultado general en (3.15) para el caso binario, resultando

$$c_{10}P_{H|\mathbf{X}}(0|\mathbf{x}) + c_{11}P_{H|\mathbf{X}}(1|\mathbf{x}) \underset{D=1}{\overset{D=0}{\gtrless}} c_{00}P_{H|\mathbf{X}}(0|\mathbf{x}) + c_{01}P_{H|\mathbf{X}}(1|\mathbf{x}) \tag{3.35}$$

Reagrupando términos, resulta

$$(c_{10}-c_{00})P_{H|\mathbf{X}}(0|\mathbf{x}) \underset{D=1}{\overset{D=0}{\gtrless}} (c_{01}-c_{11})P_{H|\mathbf{X}}(1|\mathbf{x}) \tag{3.36}$$

y, dado que $c_{10} > c_{00}$ y $c_{01} > c_{11}$, podemos escribir

$$\frac{P_{H|\mathbf{X}}(1|\mathbf{x})}{P_{H|\mathbf{X}}(0|\mathbf{x})} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{c_{10}-c_{00}}{c_{01}-c_{11}} \tag{3.37}$$

expresión que nos dice que el decisor binario de mínimo riesgo queda determinado por una función discriminante igual al cociente de probabilidades a posteriori y un umbral igual al cociente de los costes incrementales.

Aplicando el teorema de Bayes se puede obtener la regla alternativa (pero equivalente)

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{(c_{10}-c_{00})P_H(0)}{(c_{01}-c_{11})P_H(1)} \tag{3.38}$$

que nos permite expresar el decisor bayesiano utilizando como función discriminante el cociente de verosimilitudes de las hipótesis.

Decisor MAP

Análogamente, podemos particularizar las reglas para el decisor MAP

$$P_{H|\mathbf{X}}(1|\mathbf{x}) \underset{D=0}{\overset{D=1}{\gtrless}} P_{H|\mathbf{X}}(0|\mathbf{x}) \tag{3.39}$$

que podemos expresar de manera alternativa, mediante el cociente de verosimilitudes de las hipótesis, aplicando el Teorema de Bayes:

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{P_H(0)}{P_H(1)} \tag{3.40}$$

Recordemos que el decisor MAP minimiza la probabilidad de error, que en un caso binario puede escribirse como

$$P_e = r_{\phi_{\text{MAP}}} = P\{D \neq H\} = P_H(1)P_{\text{M}} + P_H(0)P_{\text{FA}} \tag{3.41}$$

3.3.4 Decisor ML

Del mismo modo que con los decisores de mínimo riesgo, y en este caso partiendo de la expresión (3.20), se puede obtener una regla de umbral para el cálculo del decisor ML en el caso binario:

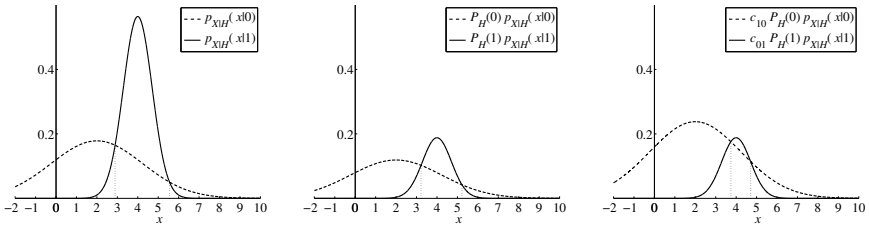
$$p_{\mathbf{X}|H}(\mathbf{x}|1) \underset{D=0}{\overset{D=1}{\gtrless}} p_{\mathbf{X}|H}(\mathbf{x}|0) \tag{3.42}$$

Example 3.10. Consideremos el problema de decisión binario unidimensional dado por las verosimilitudes gaussianas de medias $m_0 = 2$ y $m_1 = 4$ y varianzas $v_0 = 5$ y $v_1 = 0.5$, siendo las probabilidades a priori de las hipótesis $P_H(0) = 2/3$ y $P_H(1) = 1/3$. La fig. 3.4(a) muestra las verosimilitudes de ambas hipótesis. Sus puntos de corte marcan las fronteras del decisor ML, resultando $\mathcal{X}_1 = [2.89, 5.55]$ (y $\mathcal{X}_0 = \mathbb{R} - \mathcal{X}_1$). La fig. 3.4(b) muestra el decisor MAP. Dado que la hipótesis H_0 es, a priori, más probable, la región de decisión asociada a $D = 1$ se reduce con respecto a la del decisor ML y está dada por $\mathcal{X}_1 = [3.22, 5.22]$. Si, además, suponemos costes $c_{10} = 2$, $c_{01} = 1$ y $c_{00} = c_{11} = 0$, los errores al decidir $D = 1$ tienen mayor coste, y por tanto la región de decisión, $\mathcal{X}_1 = [3.74, 4.70]$, es todavía menor.

Example 3.11. Considere un problema de decisión dado por las siguientes verosimilitudes

$$\begin{aligned} p_{x|H}(x|0) &= 2\exp(-2x) \quad x > 0 \\ p_{x|H}(x|1) &= \exp(-x) \quad\;\; x > 0 \end{aligned}$$

donde se sabe que $P_H(0) = \frac{2}{3}$ y se tienen los siguientes costes: $c_{00} = c_{11} = 0$, $c_{10} = 1$ y $c_{01} = 3$. Obtenga el decisor de mínimo riesgo, el de mínima probabilidad



(a) Decisor ML (b) Decisor MAP (c) Decisor de mínimo riesgo

Fig. 3.4. Decisores para un problema de decisión binario unidimensional dado por las verosimilitudes gaussianas de medias $m_0 = 2$ y $m_1 = 4$ y varianzas $v_0 = 5$ y $v_1 = 0.5$, siendo las probabilidades a priori de las hipótesis $P_H(0) = 2/3$ y $P_H(1) = 1/3$. La figura (c) muestra el decisor de riesgo mínimo con costes $c_{10} = 2$, $c_{01} = 1$ y $c_{00} = c_{11} = 0$.

de error y el decisor ML, calculando para cada uno de ellos su probabilidad de error. Comenzamos calculando el decisor de mínimo riesgo a partir de la expresión (3.38):

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} = \frac{\exp(-x)}{2\exp(-2x)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{(c_{10}-c_{00})P_H(0)}{(c_{01}-c_{11})P_H(1)} = \frac{1-0}{3-0}\cdot\frac{2/3}{1/3}$$

$$\exp(x) \underset{D=0}{\overset{D=1}{\gtrless}} \frac{4}{3}$$

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \ln\frac{4}{3} = 0.28$$

Para el cálculo de la probabilidad de error, comenzamos calculando su P_{FA} y P_{M} mediante las expresiones (3.22) y (3.23):

$$\begin{aligned} P_{\text{FA}} = P_{D|H}(1|0) &= \int_{\mathcal{X}_1} p_{X|H}(x|0)dx = \int_{\ln\frac{4}{3}}^{\infty} 2\exp(-2x)dx \\ &= \exp\left(-2\ln\frac{4}{3}\right) = \frac{9}{16} \end{aligned}$$

$$\begin{aligned} P_{\text{M}} = P_{D|H}(0|1) &= \int_{\mathcal{X}_0} p_{X|H}(x|1)dx = \int_{0}^{\ln\frac{4}{3}} \exp(-x)dx \\ &= 1-\exp\left(\ln\frac{4}{3}\right) = \frac{1}{4} \end{aligned}$$

aplicando (3.41) tenemos la probabilidad de error

$$P_e = P_H(1)P_{\text{M}} + P_H(0)P_{\text{FA}} = \frac{1}{3}\cdot\frac{1}{4} + \frac{2}{3}\cdot\frac{9}{16} = \frac{11}{24}$$

Dado que el decisor de mínima probabilidad de error es el MAP, partimos de la expresión (3.40) para su calculo

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} = \frac{\exp(-x)}{2\exp(-2x)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{P_H(0)}{P_H(1)} = \frac{2/3}{1/3}$$

$$\exp(x) \underset{D=0}{\overset{D=1}{\gtrless}} 4$$

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \ln 4 = 1.38$$

y su probabilidad de erro la podemos calcular con:

$$P_{\text{FA}} = P_{D|H}(1|0) = \int_{\ln 4}^{\infty} 2\exp(-2x)dx = \exp\left(-2\ln 4\right) = \frac{1}{16}$$

$$P_{\text{M}} = P_{D|H}(0|1) = \int_{0}^{\ln 4} \exp(-x)dx = 1 - \exp\left(\ln 4\right) = \frac{3}{4}$$

$$P_e = P_H(1)P_{\text{M}} + P_H(0)P_{\text{FA}} = \frac{1}{3}\cdot\frac{3}{4} + \frac{2}{3}\cdot\frac{1}{16} = \frac{7}{24}$$

Por último, para el caso del decisor ML partimos de la expresión (3.42):

$$p_{\mathbf{X}|H}(\mathbf{x}|1) = \exp(-x) \underset{D=0}{\overset{D=1}{\gtrless}} p_{\mathbf{X}|H}(\mathbf{x}|0) = 2\exp(-2x)$$

$$\exp(x) \underset{D=0}{\overset{D=1}{\gtrless}} 2$$

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \ln 2 = 0.69$$

y en este caso la probabilidad de error será:

$$P_{\text{FA}} = P_{D|H}(1|0) = \int_{\ln 2}^{\infty} 2\exp(-2x)dx = \exp\left(-2\ln 2\right) = \frac{1}{4}$$

$$P_{\text{M}} = P_{D|H}(0|1) = \int_{0}^{\ln 2} \exp(-x)dx = 1 - \exp\left(\ln 2\right) = \frac{1}{2}$$

$$P_e = P_H(1)P_{\text{M}} + P_H(0)P_{\text{FA}} = \frac{1}{3}\cdot\frac{1}{2} + \frac{2}{3}\cdot\frac{1}{4} = \frac{1}{3}$$

3.3.5 Decisores no Bayesianos

Cociente de verosimilitudes

La regla de decisión dada por la ec. (3.38) se conoce como *Test de Cociente de Verosimilitudes*, o LRT (*Likelihood Ratio Test*). En su forma más general, un LRT tiene la forma

$$\frac{p_{\mathbf{X}|H}(\mathbf{x}|1)}{p_{\mathbf{X}|H}(\mathbf{x}|0)} \underset{D=0}{\overset{D=1}{\gtrless}} \eta \tag{3.43}$$

La ec. (3.43) es una interesante alternativa a (3.38) cuando las probabilidades a priori de las hipótesis son desconocidas (bien porque no se dispone de datos para estimarla o bien porque dichas probabilidades varían con el tiempo, y los datos históricos no proporcionan información fiable sobre éstas), o bien por que no se conocen los parámetros de coste (por la dificultad que, en la práctica, puede plantear la cuantificación numérica de los costes asociados a cada par decisión-hipótesis). Al depender solamente de las verosimilitudes, la ec. (3.43) evita la necesidad de considerar las hipótesis como variables aleatorias, pues, aunque son desconocidas, no es necesario calcular probabilidades sobre ellas. En teoría de la decisión, suele decirse que un decisor LRT en el que el parámetro η no depende de las probabilidades a priori es *no bayesiano*, en contraste con los decisores bayesianos, que son aquellos que modelan lo desconocido (en este caso, la hipótesis correcta) como variable aleatoria. Observe que, bajo esta perspectiva, el decisor ML es un caso particular de decisor no bayesiano dado por $\eta = 1$

Curva característica de operación

Pese a sus ventajas, el LRT tiene un parámetro libre, η, y por tanto se precisa algún procedimiento para asignarle un valor. Dado que las prestaciones de un decisor (el riesgo o la probabilidad de error) son combinación lineal de las probabilidades de falsa alarma, P_{FA} y de pérdida, P_{M} (o, equivalentemente, de detección, $P_{\text{D}} = 1-P_{\text{M}}$), resulta especialmente util representar la variación de estos dos parámetros con η. La representación de P_D en función de P_{FA} se conoce como **Curva Característica de Operación**, o ROC (*Receiver Operating Curve*).

Example 3.12. El LRT binario con verosimilitudes

$$p_{X|H}(x|1) = 2x, \qquad 0 \le x \le 1 \tag{3.44}$$

$$p_{X|H}(x|0) = 2(1-x), \qquad 0 \le x \le 1 \tag{3.45}$$

tiene la forma

$$\frac{x}{1-x} \underset{D=0}{\overset{D=1}{\gtrless}} \eta \tag{3.46}$$

o, equivalentemente

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \frac{\eta}{1+\eta} \tag{3.47}$$

Por tanto, aplicando (3.22) y (3.23)

$$P_{\text{FA}} = \int_{\mathcal{X}_1} p_{X|H}(x|0)dx = \int_{\frac{\eta}{1+\eta}}^{1} 2(1-x)dx = \frac{1}{(1+\eta)^2} \tag{3.48}$$

$$P_{\text{D}} = \int_{\mathcal{X}_1} p_{X|H}(x|1)dx = \int_{\frac{\eta}{1+\eta}}^{1} 2xdx = 1 - \frac{\eta^2}{(1+\eta)^2} \tag{3.49}$$

La ROC para este LRT se representa en la fig. 3.5.

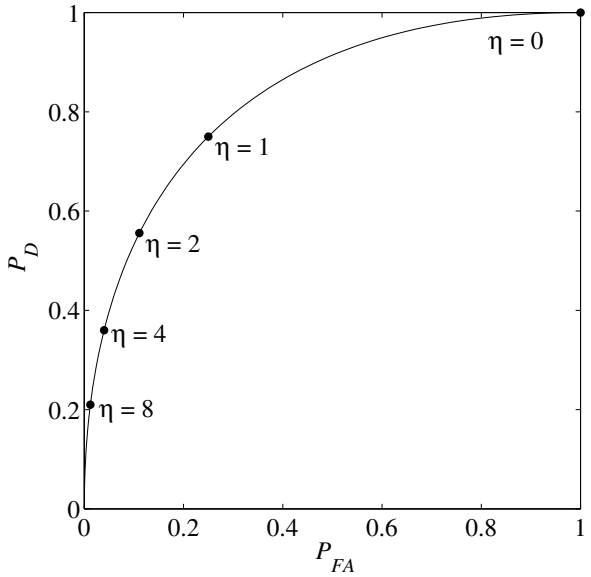


Fig. 3.5. Curva característica de operación (ROC) del LRT para las verosimilitudes del ejemplo 3.12

Es interesante observar en la fig. 3.5 del ejemplo 3.12 cómo, al asignar un valor a η, elegimos un punto de operación del decisor a lo largo de la ROC. En particular, para

$$\eta = \frac{(c_{10} - c_{00})P_H(0)}{(c_{01} - c_{11})P_H(1)}, \tag{3.50}$$

obtendremos el punto de operación del decisor bayesiano. Por tanto, los decisores Bayesianos operan en un punto de la ROC del LRT que depende de los parámetros de coste y las probabilidades a priori de las hipótesis.

El concepto de ROC no está ligado exclusivamente a los decisores tipo LRT. Cualquier familia de decisores dependiente de un parámetro tiene una curva característica de operación, que refleja las variación de las probabilidades de detección y de falsa alarma en función de dicho parámetro.

Example 3.13. La línea continua de la Figura 3.6 muestra la ROC del LRT dado por las verosimilitudes del ejemplo 3.10. Como vimos en dicho ejemplo, el decisor LRT para estas verosimilitudes es de tipo intervalo (decide 1 para todas las muestras que caen dentro de un intervalo finito, y 0 en el resto). Por tanto, los decisores *de umbral sobre* x, de la forma

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{3.51}$$

no son LRT. La ROC de la familia de decisores dados por (3.51), que recorre los puntos de operación que se obtienen variando μ, se muestra en la figura 3.6 en trazo discontinuo.

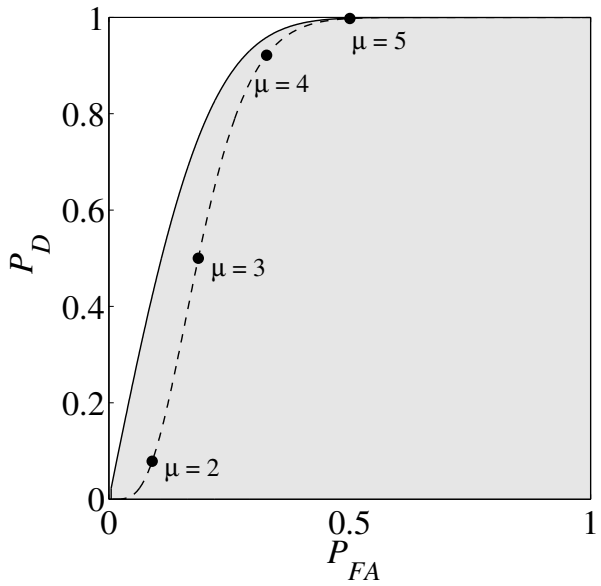


Fig. 3.6. ROC correspondiente al LRT dado por las verosimilitudes del ejemplo 3.10. La línea a trazos muestra las prestaciones de los decisores de umbral

Resulta interesante observar que, en el ejemplo anterior, la ROC del LRT está siempre por encima de la ROC de los decisores de umbral. Esto significa que, por cada punto de la ROC del decisor de umbral, existe al menos un punto de la ROC del LRT con mayor P_{D} y menor P_{FA}. Dicho de otra manera: por cada decisor de umbral, existe un decisor LRT de mejores prestaciones.

Lo anterior refleja una propiedad de caracter general: la ROC del LRT establece un límite a las prestaciones alcanzables en un problema de decisión, y no existe ningún decisor que pueda operar por encima de la ROC del LRT, [3]. Para las verosimilitudes del ejemplo 3.13, podemos afirmar que todos los decisores posibles tienen un punto de operación en la región sombreada de la Figura 3.6.

Cuando no es posible determinar el decisor bayesiano, debe utilizarse un criterio alternativo para elegir un punto de operación en la ROC. El criterio ML es una posibilidad. Mencionaremos en lo que sigue otras dos.

Detectores de Neyman-Pearson

Se conoce como detector de Neyman-Pearson (NP) a aquel que, estableciendo una cota superior a la probabilidad de falsa alarma, maximiza la probabilidad de detección. Mátematicamente, si α es la cota,

$$\phi^* = \underset{\phi}{\text{argmax}} \{P_{\text{D}}\}, \text{ sujeto a } P_{\text{FA}} \leq \alpha \tag{3.52}$$

Example 3.14. Para las verosimilitudes del ejemplo 3.12, podemos diseñar el decisor NP dado por la cota $\alpha = 0.1$. La región sombreada de la figura 3.7 muestra el conjunto de puntos de operación que satisfacen esta cota. Se observa que el valor máximo de la probabilidad de detección se obtiene tomando el valor de P_{D} que está sobre la ROC exactamente para $P_{\text{FA}} = \alpha$.

Igualando $P_{\text{FA}} = \alpha$ en (3.48) y despejando η, resulta

$$\eta = \frac{1}{\sqrt{P_{\text{FA}}}} - 1 \approx 2.1623 \tag{3.53}$$

y, aplicando (3.49) se obtiene $P_D = 0.53$. Por tanto, el decisor de Neyman-Pearson opera en el punto $(0.01, 0.53)$.

El ejemplo 3.14 ilustra un procedimiento general para determinar el decisor LRT de Neyman-Pearson cuando la ROC es una curva continua[4]: Se toma $P_{\text{FA}} = \alpha$ y se

[3] La demostración de esta propiedad no es difícil, y se basa en la relación (3.50), que establece que cada punto del LRT, $(P_{\text{FA}}(\eta), P_{\text{D}}(\eta))$, es punto de operación de algún decisor Bayesiano (por ejemplo, el decisor LRT de umbral η es también, de acuerdo con (3.50) el decisor Bayesiano para $P_H(0) = P_H(1)$, $c_{10} = \eta$, $c_{01} = 1$, $c_{00} = c_{11} = 0$). Por tanto, el decisor que opera en el punto $(P_{\text{FA}}(\eta), P_{\text{D}}(\eta))$ minimiza un coste medio. Ahora bien, cualquier punto de operación $(P^*_{\text{FA}}, P^*_{\text{D}})$ por encima y a la izquierda de $(P_{\text{FA}}(\eta), P_{\text{D}}(\eta))$ (es decir, con $P_{\text{FA}}(\eta) \leq P^*_{\text{FA}}$ y $P_{\text{D}}(\eta) \geq P^*_{\text{D}}$) tendría menor coste medio. Por lo tanto, ningún decisor puede operar en $(P^*_{\text{FA}}, P^*_{\text{D}})$

[4] La ROC del LRT puede ser discontinua, por ejemplo cuando las observaciones son variables aleatorias discretas. No estudiaremos estos casos aquí.

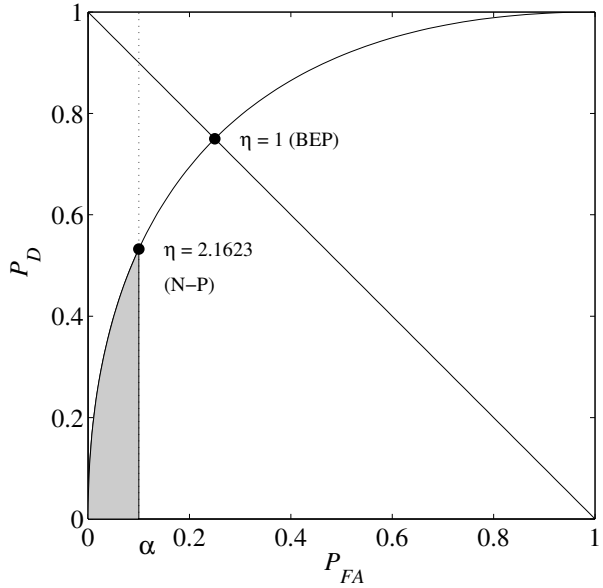


Fig. 3.7. Punto de operación del decisor NP y *Break Even Point* (BEP) para el LRT del ejemplo 3.12

calcula el valor de η en la ROC que proporciona dicha P_{FA}. Con frecuencia, el calculo de η no puede hacerse analíticamente, y hay que recurrir a métodos numéricos.

Cuando $p_{\mathbf{X}|H}(\mathbf{x}|0)$ es una distribución de tipo exponencial, resulta conveniente utilizar el logaritmo del cociente de verosimilitudes como función discriminante

$$\Lambda = g((\mathbf{X}) = \log\left(\frac{p_{\mathbf{X}|H}(\mathbf{X}|1)}{p_{\mathbf{X}|H}(\mathbf{X}|0)}\right) \tag{3.54}$$

de tal modo que el LRT puede escribirse como

$$\begin{array}{c} D=1 \\ \Lambda \gtrless \eta' \\ D=0 \end{array} \tag{3.55}$$

siendo $\eta' = \log(\eta)$. La probabilidad de falsa alarma será, entonces

$$P_{\text{FA}} = \int_{\eta'}^{\infty} P_{\Lambda|H}(\lambda|0)d\lambda \tag{3.56}$$

Tomando $P_{\text{FA}} = \alpha$ en (3.56) y despejando η', se obtiene el LRT.

Decisores minimax

Otra forma alternativa de elegir el punto de operación consiste en otorgar la misma importancia a los errores de pérdida que a los de falsa alarma, es decir, elegir η de tal

modo que $P_{\text{M}} = P_{\text{FA}}$ o, equivalentemente

$$P_{\text{D}} = 1 - P_{\text{FA}} \tag{3.57}$$

Gráficamente dicho punto es la intersección de la ROC con la recta $P_{\text{D}} = 1 - P_{\text{FA}}$, y se denomina *Break Even Point* (BEP).

Example 3.15. Para las verosimilitudes del ejemplo 3.12, igualando $P_{\text{FA}} = 1 - P_D$ en (3.48) y (3.49), y despejando, se obtiene $\eta = 1$, tal y como se muestra en la figura 3.7.

El BEP tiene una interesante propiedad: aplicando (3.41) se comprueba que

$$P_e = P_{\text{FA}} = P_{\text{M}} \tag{3.58}$$

y, por tanto, la probabilidad de error no depende de las probabilidades a priori. Esto hace al decisor que opera en el BEP muy *robusto* frente a cambios en las probabilidades a priori. Para cualquier otro decisor LRT de parámetro η, la probabilidad de error (3.41) varía linealmente con las probabilidades a priori. En el *caso peor*, resulta

$$\max_{P_H(0), P_H(1)} \{Pe\} = \max\{P_{\text{FA}}(\eta), P_{\text{M}}(\eta)\} \tag{3.59}$$

Puede demostrarse, de hecho, que el punto BEP minimiza la probabilidad de error *caso peor*,

$$\eta^{\text{BEP}} = \operatorname*{argmin}_{\eta} \max\{P_{\text{FA}}(\eta), P_{\text{M}}(\eta)\} \tag{3.60}$$

Por este motivo, el decisor que opera en el BEP se denomina decisor *minimax*.

3.4 El caso Gaussiano

Supongamos que

$$p_{\mathbf{X}|H}(\mathbf{x}|i) = \frac{1}{(2\pi)^{N/2}|\mathbf{V}_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^T\mathbf{V}_i^{-1}(\mathbf{x}-\mathbf{m}_i)\right) \tag{3.61}$$

en tal caso, aplicando logaritmos sobre el decisor LRT en (3.38), podemos escribir

$$\begin{aligned} &-\frac{1}{2}\log|\mathbf{V}_1| - \frac{1}{2}(\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1) \\ &+\frac{1}{2}\log|\mathbf{V}_0| + \frac{1}{2}(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0) \underset{D=0}{\overset{D=1}{\gtrless}} \log(\eta) \end{aligned} \tag{3.62}$$

y, agrupando términos

$$(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}_0^{-1}(\mathbf{x}-\mathbf{m}_0) - (\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}_1^{-1}(\mathbf{x}-\mathbf{m}_1) \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{3.63}$$

siendo

$$\mu = 2\log(\eta) + \log|\mathbf{V}_1| - \log|\mathbf{V}_0| \tag{3.64}$$

La ec. (3.63) muestra que el decisor óptimo en un caso gaussiano está dado por una regla de decisión que es un polinomio de segundo grado. Las fronteras de decisión resultantes son, por tanto, superficies cuádricas. En particular, en problemas bidimensionales, resultan hipérbolas, parábolas, elipses o líneas rectas.

Example 3.16. La figura 3.8 muestra las frontera de decisión ML para un problema binario bidimensional gausiano con variables

$$\mathbf{X}|0 \sim G\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1.2 & 0.43\\0.43 & 1.75\end{pmatrix}\right) \tag{3.65}$$

$$\mathbf{X}|1 \sim G\left(\begin{pmatrix}3\\3\end{pmatrix}, \begin{pmatrix}2 & 0\\0 & 1\end{pmatrix}\right) \tag{3.66}$$

donde $\mathbf{X}|i \sim G$ indica que $p_{\mathbf{X}|H}(\mathbf{x}|0)$ es la función G, y $G(\mathbf{m}, \mathbf{V})$ denota la densidad de probabilidad gausiana de media $\mathbf{m}$ y matriz de varianzas $\mathbf{V}$.

En tonos grises se muestra la distribución de $p_{\mathbf{X}}(\mathbf{x})$ siendo más oscuras las zonas de mayor densidad. Las líneas blancas muestran las curvas de nivel de cada una de las verosimilitudes. La frontera de decisión es una hipérbola con dos ramas, aunque solamente una de ellas es visible (la otra recorre valores de x_2 superiores a los que abarca la figura).

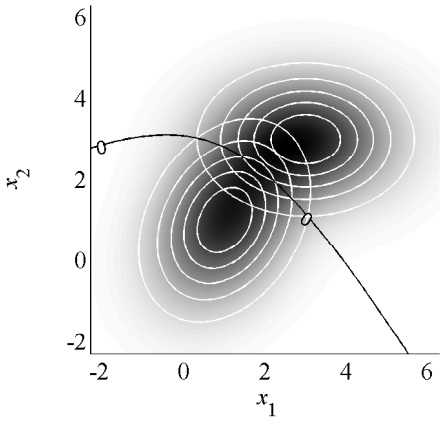


Fig. 3.8. Frontera del decisor ML para un problema de decisión binario bidimensional. En tonos grises se muestra la distribución de $p_{\mathbf{X}}(\mathbf{x})$ siendo más oscuras las zonas de mayor densidad. Las líneas blancas muestran las curvas de nivel de cada una de las verosimilitudes.

Example 3.17. La figura 3.9 muestra la frontera de decisión ML para un problema binario bidimensional gausiano con variables

$$\mathbf{X}|0 \sim G\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.7 & 0\\0 & 0.7\end{pmatrix}\right) \tag{3.67}$$

$$\mathbf{X}|1 \sim G\left(\begin{pmatrix}0.2\\0.4\end{pmatrix}, \begin{pmatrix}0.5 & 0\\0 & 0.2\end{pmatrix}\right) \tag{3.68}$$

En este caso la frontera de decisión es una elipse.

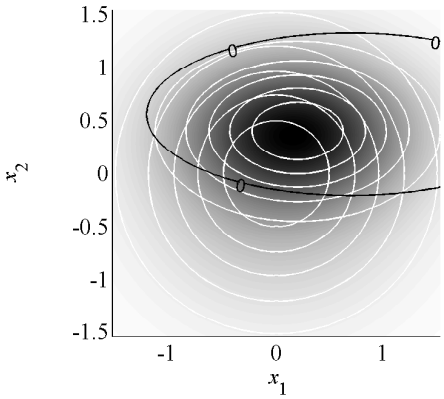


Fig. 3.9. Ejemplo de un problema de decisión binario bidimensional gausiano con frontera elíptica.

Hay algunos casos particulares de interés:

3.4.1 Varianzas iguales

En tal caso, (3.63) se reduce a

$$(\mathbf{x}-\mathbf{m}_0)^T\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}_0) - (\mathbf{x}-\mathbf{m}_1)^T\mathbf{V}^{-1}(\mathbf{x}-\mathbf{m}_1) \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{3.69}$$

Descomponiendo la forma cuadrática, y observando que el término $\mathbf{x}^T\mathbf{V}^{-1}\mathbf{x}$ aparece dos veces con sentidos opuestos, y se cancela, resulta

$$2(\mathbf{m}_1-\mathbf{m}_0)^T\mathbf{V}^{-1}\mathbf{x} + \mathbf{m}_0^T\mathbf{V}^{-1}\mathbf{m}_0 - \mathbf{m}_1^T\mathbf{V}^{-1}\mathbf{m}_1 \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{3.70}$$

que es un decisor lineal. Obsérvese que, además, en tal caso $\mu = 2\log(\eta)$, y es independiente de la varianza de las distribuciones.

Example 3.18. La figura 3.10 muestra las fronteras de decisión para un problema binario bidimensional gausiano con variables

$$\mathbf{X}|0 \sim G\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}0.44 & 0.32\\0.32 & 0.81\end{pmatrix}\right) \tag{3.71}$$

$$\mathbf{X}|1 \sim G\left(\begin{pmatrix}3\\3\end{pmatrix}, \begin{pmatrix}0.44 & 0.32\\0.32 & 0.81\end{pmatrix}\right) \tag{3.72}$$

El número etiquetando cada frontera indica el valor de $\log(\eta)$, de tal manera que $\log(\eta) = 0$ se corresponde con el decisor ML.

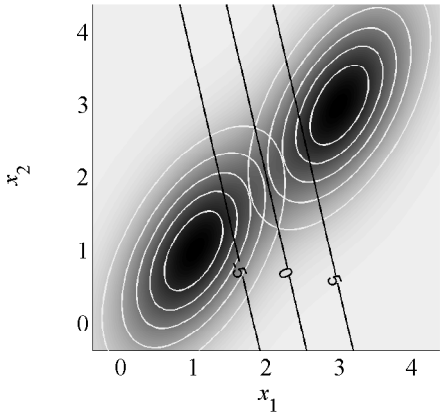


Fig. 3.10. Decisores para un problema de decisión binario bidimensional con matrices de varianzas iguales.

3.4.2 Medias nulas

Suponiendo $\mathbf{m}_0 = \mathbf{m}_1 = 0$, resulta

$$\mathbf{x}^T \left(\mathbf{V}_0^{-1} - \mathbf{V}_1^{-1}\right) \mathbf{x} \underset{D=0}{\overset{D=1}{\gtrless}} \mu \tag{3.73}$$

Example 3.19. La figura 3.10 muestra la frontera de decisión ML para un problema binario bidimensional gausiano con variables

$$\mathbf{X}|0 \sim G\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.62 & -0.22\\-0.22 & 0.37\end{pmatrix}\right) \tag{3.74}$$

$$\mathbf{X}|1 \sim G\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 2\end{pmatrix}\right) \tag{3.75}$$

La región $\mathcal{X}_0$ se corresponde con el interior de la elipse. Al ser mayor la varianza de $\mathbf{X}|1$ en todas direcciones, lejos del origen de coordenadas el decisor siempre toma $D = 1$.

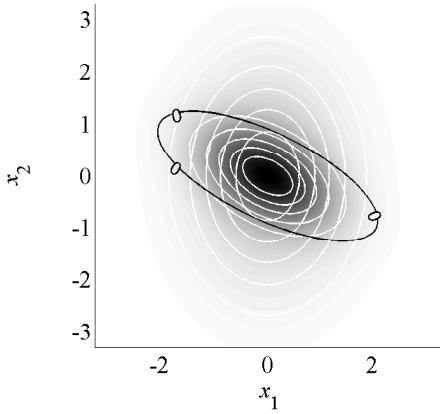


Fig. 3.11. Un problema de decisión binario bidimensional gausiano con medias nulas y frontera elíptica.

Example 3.20. La figura 3.10 muestra la frontera de decisión ML para un problema binario bidimensional gausiano con variables

$$\mathbf{X}|0 \sim G\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.33 & 0.39\\0.39 & 0.77\end{pmatrix}\right) \tag{3.76}$$

$$\mathbf{X}|1 \sim G\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}0.39 & -0.19\\-0.19 & 0.16\end{pmatrix}\right) \tag{3.77}$$

En este ejemplo, la varianza de $\mathbf{X}|1$ es dominante solamente en algunas direcciones, pero no en otras. Como resultado, la frontera de decisión es una hipérbola.

3.5 Apéndices

3.5.1 Diseño analítico de decisores con costes dependientes de la observación

En un caso general la función de coste que modela las consecuencias derivadas de cada decisión no tiene porque ser independiente de la observación y es necesario definir el coste como una función de $\mathbf{x}$. En este anexo consideraremos esta situación y derivaremos las reglas de decisión resultantes de considerar, en primer lugar, funciones de costes dependientes de la decisión y la observación y, en segundo lugar, dependientes de la decisión, la hipótesis y el valor de la observación.

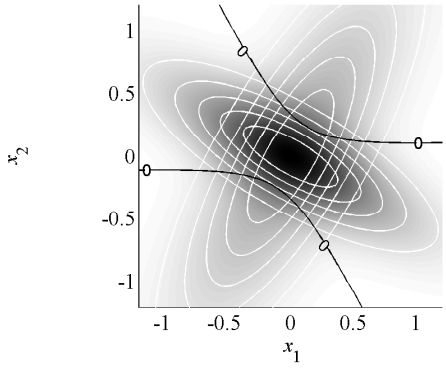


Fig. 3.12. Un problema de decisión binario bidimensional gausiano con frontera hiperbólica.

Costes dependientes de la decisión y la observación

En este primer caso se considerará que las consecuencias derivadas de decidir $D = d$ cuando la observación es $\mathbf{X} = \mathbf{x}$ vienen modeladas por una función de coste $c(d, \mathbf{x}) \in \mathbb{R}$. Entonces el **riesgo** del decisor dado por la función de decisión ϕ pueden medirse por su coste medio:

$$\begin{aligned} r_\phi &= \mathbb{E}\{c(D, \mathbf{X})\} = \mathbb{E}\{c(\phi(\mathbf{X}), \mathbf{X})\} \\ &= \int_{\mathcal{X}} c(\phi(\mathbf{x}), \mathbf{x}) p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \end{aligned} \tag{3.78}$$

La ecuación anterior nos proporciona de forma inmediata una regla para diseñar el decisor óptimo: la integral será mínima cuando sean mínimos los valores del integrando. Por tanto, el decisor óptimo será aquel que, para cada valor de $\mathbf{x}$, tome la decisión

$$\phi^*(\mathbf{x}) = \operatorname*{argmin}_d c(d, \mathbf{x}) \tag{3.79}$$

Example 3.21. Pedro desea comprar un nuevo modelo de reloj. En la relojería TIC, próxima a su casa lo venden por 130 Euros, pero sabe que en la relojería TAC puede conseguirlo por 124 euros. Es más barato, pero la tienda está a 40 Km, y tendría que ir a buscarlo en su coche, que consume 5 litros cada 100 Km. Tras observar el precio de la gasolina, x, Pedro debe decidir entre comprar el reloj en TIC ($D = 0$) o bien en TAC ($D = 1$). Midiendo el coste de cada decisión en términos exclusivamente económicos, resultan los costes:

$$c(0, x) = 130 \tag{3.80}$$

$$c(1, x) = 124 + 40 \cdot 0.05 \cdot x \tag{3.81}$$

La regla decisión óptima dado x será

$$\begin{array}{c} D=0 \\ x \gtrless 3 \\ D=1 \end{array} \tag{3.82}$$

Es decir, solamente si el precio de la gasolina sube por encima de los 3 Euros por litro, compensa comprar el reloj en la tienda más próxima.

Desafortunadamente, la aplicación práctica de la regla (3.79) tropieza con una dificultad fundamental: la premisa de partida (a saber, que es posible cuantificar las consecuencias derivadas de cada decisión) es, en general, falsa. Las consecuencias derivadas de cada acción tienen una componente impredecible antes de tomar la decisión, y deben modelarse probabilísticamente.

Example 3.22. Pedro se da cuenta de que su estimación del consumo de gasolina en su coche es poco realista. El consumo real depende mucho de las condiciones del tráfico. Considera más adecuado modelar el consumo de su coche como una variable aleatoria uniformemente distribuida entre 4 y 8 ($U(4,8)$). En tal caso, resultan los costes:

$$c(0,x) = 130 \tag{3.83}$$

$$C(1,x) = 124 + 0.4Gx \tag{3.84}$$

donde G es una variable aleatoria $U(4,8)$. Observe que $c(0,x)$ se denota con minúsculas, porque no tiene ninguna componente aleatoria, mientras que $C(1,x)$ sí la tiene.

De forma general, debemos considerar el coste $C(d,\mathbf{x})$ como variable aleatoria. Aplicando la fórmula de la esperanza matemática total, el coste medio global del decisor ϕ será

$$r_\phi = \mathbb{E}\{C(\phi(\mathbf{X}),\mathbf{X})\} = \int_{\mathcal{X}} \mathbb{E}\{C(\phi(\mathbf{X}),\mathbf{X})|\mathbf{X}=\mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{3.85}$$

Se puede obtener una expresión alternativa descomponiendo la integral por regiones de decisión, que facilita el cálculo

$$r_\phi = \sum_{d=0}^{M-1} \int_{\mathcal{X}_d} \mathbb{E}\{C(d,\mathbf{X})|\mathbf{X}=\mathbf{x}\} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \tag{3.86}$$

En todo caso, partiendo de (3.85), resulta evidente que la decisión óptima para una observación $\mathbf{x}$ será

$$\phi^*(\mathbf{x}) = \operatorname*{argmin}_d \mathbb{E}\{C(d,\mathbf{X})|\mathbf{X}=\mathbf{x}\} \tag{3.87}$$

Es decir, el decisor óptimo será aquel que, para cada observación, tome la decisión cuyo *riesgo condicional* sea mínimo.

Example 3.23. Continuando el ejemplo 3.22, los costes medios dada la observación x serán

$$\mathbb{E}\{c(0,x)|x\} = 130 \tag{3.88}$$
$$\mathbb{E}\{C(1,x)|x\} = 124 + 0.4\mathbb{E}\{G\}x = 124 + 2.4x \tag{3.89}$$

de donde resulta el decisor óptimo

$$\begin{array}{c} D=0 \\ x \gtrless 2.5 \\ D=1 \end{array} \tag{3.90}$$

Costes dependientes de la decisión, la hipótesis y la observación

Aunque (3.87) proporciona una solución teórica al diseño de un decisor, la aplicación práctica de esta ecuación pasa por determinar un procedimiento para calcular el riesgo condicional, $\mathbb{E}\{C(d,\mathbf{X})|\mathbf{X}=\mathbf{x}\}$, lo que no siempre es fácil. Una estrategia frecuente consiste en analizar las posibles circunstancias o **hipótesis**, no observadas por el decisor, bajo las que puede tener lugar la decisión, y determinar las consecuencias (costes) de cada decisión bajo cada una de dichas circunstancias.

Example 3.24. Pedro sabe que el consumo de su coche es muy diferente dependiendo de si hay poco tráfico (hipótesis $H=0$) o mucho (hipótesis $H=1$). Para $H=0$, el consumo G en la ec. (3.84) sigue una distribución $p_{G|H}(g|0) = 4(8-g)$, con $4 < g \leq 8$. Por el contrario, bajo hipótesis $H=1$, $p_{G|H}(g|1) = 4(g-8)$, con $4 < g \leq 8$. Si Pedro conociera que hay poco tráfico, podría utilizar los costes condicionales

$$\mathbb{E}\{c(0,x)|x, H=0\} = 130 \tag{3.91}$$
$$\mathbb{E}\{C(1,x)|x, H=0\} = 124 + 0.4\mathbb{E}\{G|H=0\}x = 124 + \frac{20}{3}x \tag{3.92}$$

que conducen a la regla

$$\begin{array}{c} D=0 \\ x \gtrless 2.5 \\ D=1 \end{array} \tag{3.93}$$

Por el contrario, si hay mucho tráfico, el umbral de decisión óptima asciende a 2.8125. Por tanto, en el rango $2.5 \leq x \leq 2.8125$, la mejor decisión varía dependiendo de cuál de las dos hipótesis es correcta.

En el ejemplo anterior, se comprueba que cada hipótesis conduciría a una regla de decisión distinta.

3.6 Problemas

3.1. Considere el problema de decisión con tres hipótesis dado por la observación $\mathbf{x} = (x_1, x_2) \in [0,1]^2$ y verosimilitudes

$$p_{\mathbf{X}|H}(\mathbf{x}|0) = 2(1 - x_1) \tag{3.94}$$

$$p_{\mathbf{X}|H}(\mathbf{x}|1) = 2x_1 \tag{3.95}$$

$$p_{\mathbf{X}|H}(\mathbf{x}|2) = 2x_2 \tag{3.96}$$

a) Determine el decisor ML
b) Represente las regiones de decisión.

3.2. Considere el problema de decisión dado por la observación $x \in [0,1]$, verosimilitudes

$$p_{X|H}(x|0) = 2(1 - x) \tag{3.97}$$

$$p_{X|H}(x|1) = 1 \tag{3.98}$$

y probabilidad a priori $P_H(1) = 1/4$.

a) Determine el decisor ML
b) Determine el decisor MAP.
c) Sabiendo que $c_{01} = 2$, $c_{10} = 1$, $c_{11} = c_{00} = 0$, determine el decisor de mínimo coste medio.
d) Considere un decisor de umbral sobre x de la forma

$$\begin{array}{c} D = 1 \\ x \quad \gtrless \quad \eta \\ D = 0 \end{array} \tag{3.99}$$

Calcule las probabilidades de falsa alarma, pérdida y error, en función de η.
e) Aplique el resultado a los tres decisores anteriores. Compruebe que el decisor MAP obtiene la mímima probabilidad de error.
f) Determine el riesgo para los tres decisores anteriores, utilizando los parámetros de coste del apartado (c), y compruebe que el decisor obtenido en dicho apartado obtiene el menor coste medio.

3.3. Un problema de decisión binaria unidimensional con hipótesis equiprobables obedece a las verosimilitudes

$$\begin{aligned} p_{X|H}(x|0) &= \frac{1}{6}, \quad |x| \leq 3, \\ p_{X|H}(x|1) &= \frac{3}{2}x^2, \quad |x| \leq 1 \end{aligned}$$

a) Determine el decisor ML.
b) Determine los valores de P_{FA}, P_{M} y P_{e} del decisor anterior.

3.4. Considere el problema de decisión dado por la observación $x \geq 0$ y verosimilitudes

$$p_{X|H}(x|0) = \exp(-x); \tag{3.100}$$

$$p_{X|H}(x|1) = 2\exp(-2x); \tag{3.101}$$

a) Determine el decisor LRT
b) Determine la ROC

3.5. Considere un problema de decisión binaria donde la verosimilitud de la hipótesis $H = 0$ es uniforme en el intervalo $0 < x < 1$, mientras que $p_{X|H}(x|1) = 2x$, $0 < x < 1$.

a) Obtenga la expresión genérica de un test de razón de verosimilitudes con parámetro η. Represente gráficamente sobre unos mismos ejes ambas verosimilitudes, indicando sobre dicha figura las regiones de decisión correspondientes al caso ML.
b) Obtenga la expresión analítica de la curva ROC del decisor LRT. Represente dicha curva indicando sobre la misma los puntos de operación correspondientes al decisor ML y al decisor de Neyman-Pearson con parámetro $\alpha = 0.1$ (i.e., $P_{\text{FA}} \leq \alpha = 0.1$).

3.6. La empresa E fabrica diariamente 10000 unidades de un producto. Se ha estimado que:

- La venta de una unidad en buenas condiciones reporta un beneficio neto de 3 €
- La puesta en el mercado de una unidad defectuosa ocasiona (en media) unas pérdidas de 81 €
- La retirada de una unidad (tenga o no defectos) supone pérdidas de 1 €

Se dispone de un sistema automático de inspección que obtiene, por cada unidad, una observación x_1. Se sabe que, llamando $H = 0$ a la hipótesis "la unidad no es defectuosa" y $H = 1$ a la hipótesis "la unidad es defectuosa",

$$p_{X_1|H}(x_1|0) = \exp(-x_1)u(x_1) \tag{3.102}$$

$$p_{X_1|H}(x_1|1) = \lambda_1 \exp(-\lambda_1 x_1)u(x_1) \tag{3.103}$$

siendo $\lambda_1 = 1/2$. Asimismo, se sabe también que la cadena de montaje produce, en media, una unidad defectuosa por cada 100 no defectuosas.

Se pretende incorporar un mecanismo de retirada automática de unidades defectuosas basado en la observación de x_1.

a) Diséñese el detector que proporcione a E el mayor beneficio esperado.
b) Determínese el máximo beneficio esperado (diario) que se puede obtener.
c) Una empresa ofrece a E un innovador dispositivo de inspección que proporciona, para cada producto, además de x_1, una nueva observación x_2, estadísticamente independiente de x_1, tal que

$$p_{X_2|H}(x_2|0) = \exp(-x_2)u(x_2) \tag{3.104}$$

$$p_{X_2|H}(x_2|1) = \lambda_2 \exp(-\lambda_2 x_2) u(x_2) \tag{3.105}$$

siendo $\lambda_2 = 1/4$. El coste de dicha máquina es de 6000 Euros. Determínese una expresión para el tiempo medio que tardaría E en amortizar dicha máquina.

3.7. (*) En la sección 3.4 hemos visto problemas de decision bidimensionales gaussianos caracterizados por fronteras de decisión hiperbólicas, elípticas y lineales. Dado que la parábola también es una superficie cuádrica, cabe esperar que existan valores de $\mathbf{m}_0$, $\mathbf{m}_1$, $\mathbf{V}_0$ y $\mathbf{V}_1$ que den lugar a fronteras parabólicas. Especifique un problema de decisión gaussiano que de lugar a fronteras parabólicas.

3.8. [5] En un sistema de comunicación binario el transmisor puede transmitir la señal $s_0 = 0$ o la señal $s_1 = 1$. El canal de comunicación añade, independientemente, ruido N gausiano de media nula y varianza 1, siendo la señal recibida:

$$X = s_i + N \quad i = 0, 1$$

En un momento dado el receptor observa la señal $x = 0.6$:

a) Determine por máxima verosimilitud que señal envió el transmisor. Obtenga su probabilidad de falsa alarma y de pérdida
b) Sabiendo el símbolo s_0 tiene una probabilidad a priori de ser transmitido de $2/3$, aplique el decisor MAP para decidir que señal envió el transmisor y obtenga su probabilidad de falsa alarma y de pérdida.
c) Aplique el criterio de Neyman-Pearson para decidir que señal envió el transmisor considerando que la P_{FA} debe ser menor de 0.25. Obtenga su probabilidad de pérdida.

Considere ahora que el receptor puede disponer de K observaciones independientes (todas ellas correspondientes a un único símbolo transmitido) para decidir que símbolo fue transmitido:

d) Obtenga la expresión del test de máxima verosimilitud.
d) Calcule la probabilidad de falsa alarma y de pérdida cuando se dispone de $K = 5$ y $K = 10$ observaciones.

[5] Este problema ha sido adaptado de: *H. Hsu, " Schaum's Outline of Probability, Random Variables, and Random Processes", 2nd ed., Mc Graw Hill, 2010.*

4

Decisión máquina

4.1 Diseño de clasificadores bajo enfoque máquina

Toda la teoría de la decisión estadística (basada en hipótesis) se fundamenta en el supuesto de que se dispone de un modelo estadístico de la dependencia entre hipótesis y observaciones. En la práctica, este supuesto raras veces se cumple, y el diseño del clasificador debe abordarse a partir de colecciones de datos representativos del problema.

En lo que sigue, abandonaremos la formulación general, basada en costes, del problema de decisión, y nos centraremos en el criterio de mínima probabilidad de error (es decir el caso particular con costes $c_{01} = c_{10} = 1$, $c_{00} = c_{11} = 0$). En primera aproximación, plantearemos el problema de diseño bajo enfoque máquina del modo siguiente: disponiendo de un conjunto $\mathcal{S} = \{\mathbf{x}^{(k)}, h^{(k)}, k = 1, \ldots, K\}$ de realizaciones independientes de cierta distribución $p_{X,H}(\mathbf{x}, h)$ desconocida, se pretende diseñar un clasificador $D = \phi(\mathbf{X})$ de mínima probabilidad de error.

Siendo desconocida la distribución de los datos, el diseño exacto (es decir, analítico) del clasificador óptimo (esto es, el decisor MAP) no es posible, y por lo tanto cualquier procedimiento de diseño proporcionará, en general, un decisor aproximado. Interesan, en todo caso, procedimientos que proporcionen buenas aproximaciones para K finito, y diseños asintóticamente óptimos para K infinito (es decir, que a medida que el número de observaciones disponibles para el diseño aumente, el clasificador resultante se aproxime al decisor MAP).

La evaluación de las prestaciones de un clasificador plantea problemas similares al caso de estimación máquina: el diseño del clasificador debe garantizar buenas propiedades de generalización, es decir, el clasificador debe conservar buenas prestaciones con datos no utilizados durante el entrenamiento, y para ello, la partición de los datos en conjuntos de entrenamiento ($\mathcal{S}_e$), test ($\mathcal{S}_t$) y (posiblemente) validación ($\mathcal{S}_v$) es imprescindible para evaluar las prestaciones del diseño obtenido.

Los resultados de la teoría de la decisión analítica sugieren tres estrategias diferentes para diseñar decisores óptimos a partir de $\mathcal{S}$:

1. Estimar las probabilidades a priori de las hipótesis y sus verosimilitudes, y luego aplicar (3.14)
2. Estimar las probabilidades a posteriori de las hipótesis para aplicar la regla (3.19)
3. Determinar una función discriminante de mínima probabilidad de error.

Las estrategias 1 y 2 corresponden con lo que se denomina habitualmente procedimientos semianalíticos. Los métodos máquina propiamente dichos se corresponden con la estrategia 3, que no presupone ningún modelo estadístico para las observaciones.

Discutiremos brevemente la estrategia 1, aunque el procedimiento a seguir puede extenderse fácilmente a la estrategia 2.

4.1.1 Estimación paramétrica ML para clasificación

El procedimiento (paramétrico) habitual para diseñar un clasificador mediante las estrategias 1 y 2 consiste en postular un modelo $\hat{p}_{\mathbf{X},H}(\mathbf{x}, h|\mathbf{v})$ para la verosimilitud de cada hipótesis, y estimar su parámetros, $\mathbf{v}$ mediante máxima verosimilitud sobre los datos, es decir

$$\begin{aligned}\mathbf{v}^* &= \arg\max_{\mathbf{v}} \prod_{k=1}^{K} \hat{p}_{\mathbf{X},H|\mathbf{v}}(\mathbf{x}^{(k)}, h^{(k)}|\mathbf{v}) \\ &= \arg\max_{\mathbf{v}} \sum_{k=1}^{K} \log \hat{p}_{\mathbf{X},H|\mathbf{v}}(\mathbf{x}^{(k)}, h^{(k)}|\mathbf{v}) \end{aligned} \tag{4.1}$$

o, equivalentemente, descomponiendo distribuciones conjuntas en productos de verosimilitudes por probabilidades a priori,

$$\mathbf{v}^* = \arg\max_{\mathbf{v}} \left(\sum_{k=1}^{K} \log \hat{P}_{H|\mathbf{v}}(h^{(k)}|\mathbf{v}) + \sum_{k=1}^{K} \log \hat{p}_{\mathbf{X}|H,\mathbf{v}}(\mathbf{x}^{(k)}|h^{(k)}, \mathbf{v}) \right) \tag{4.2}$$

Si suponemos que el vector de parámetros puede descomponerse como $\mathbf{v} = (\mathbf{q}, \mathbf{w})$, siendo $\mathbf{q}$ parámetros de las probabilidades a priori, y $\mathbf{w}$ parámetros de las verosimilitudes, podemos escribir

$$(\mathbf{q}^*, \mathbf{w}^*) = \arg\max_{\mathbf{q},\mathbf{w}} \left(\sum_{k=1}^{K} \log \hat{P}_{H|\mathbf{q}}(h^{(k)}|\mathbf{q}) + \sum_{k=1}^{K} \log \hat{p}_{\mathbf{X}|H,\mathbf{w}}(\mathbf{x}^{(k)}|h^{(k)}, \mathbf{w}) \right) \tag{4.3}$$

y, por tanto, podemos separar la estimación de parámetros de probabilidades a priori y verosimilitudes,

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} \sum_{k=1}^{K} \log \hat{P}_{H|\mathbf{q}}(h^{(k)}|\mathbf{q}) \tag{4.4}$$

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{k=1}^{K} \log \hat{p}_{\mathbf{X}|H,\mathbf{w}}(\mathbf{x}^{(k)}|h^{(k)}, \mathbf{w}) \tag{4.5}$$

El decisor resultante aplicará la regla

$$d = \phi(\mathbf{x}) = \arg\max_{h} \hat{P}_{H|\mathbf{q}}(h|\mathbf{q}^*)\hat{p}_{\mathbf{X}|H}(\mathbf{x}|h, \hat{\mathbf{w}}^*) \tag{4.6}$$

Estimación de probabilidades a priori

Si el número de hípótesis no es muy alto, podemos utilizar un parámetro para la probabilidad a priori de cada una de las clases, de modo que q_h defina la probabilidad a priori de la hipótesis h; es decir,

$$\hat{P}_{H|\mathbf{q}}(h|\mathbf{q}) = q_h \tag{4.7}$$

Sustituyendo (4.7) en (4.4), se obtiene

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} \sum_{k=1}^{K} \log q_{h^{(k)}} \tag{4.8}$$

Si llamamos K_h al número de muestras en $\mathcal{S}$ de la categoría h, podemos agrupar los sumandos en (4.8) por categorías, resultando

$$\mathbf{q}^* = \arg\max_{\mathbf{q}} \sum_{h=1}^{M} K_h \log(q_h) \tag{4.9}$$

El máximo (con la restricción de que $\sum_h q_h = 1$) se obtiene cuando

$$q_h^* = \frac{K_h}{K} \tag{4.10}$$

En definitiva, las probabilidades a priori se estiman como la porción de muestras de cada hipótesis en el conjunto de entrenamiento.

Estimación de verosimilitudes

Las verosimilitudes de las hipótesis habrán de estimarse utilizando (4.5).

Example 4.1. Considere el clasificador binario dado por los modelos de verosimilitudes

$$\hat{p}_{X|H}(x|0) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-m_0)^2}{2v}\right) \tag{4.11}$$

$$\hat{p}_{X|H}(x|1) = \frac{1}{\sqrt{2\pi v}} \exp\left(-\frac{(x-m_1)^2}{2v}\right) \tag{4.12}$$

Para aplicar (4.5) al vector de parámetros $\mathbf{w} = (m_0, m_1, v)$, dividiremos el sumatorio sobre todas las muestras en dos sumas, una por cada categoría. Llamando $\mathcal{I}_0$ al conjunto de índices de todas las muestras de la categoría 0 (es decir, al conjunto de valores de k para los que $h^{(k)} = 0$), y $\mathcal{I}_1$ al resto de índices, podemos re-escribir la ec. (4.5) como

$$(m_0^*, m_1^*, v^*) = \arg \max_{m_0, m_1, v} L(m_0, m_1, v) \tag{4.13}$$

siendo

$$\begin{aligned} L(m_0, m_1, v) &= \sum_{k \in \mathcal{I}_0} \log \hat{p}_{\mathbf{X}|H,\mathbf{w}}(\mathbf{x}^{(k)}|0, \mathbf{w}) + \sum_{k \in \mathcal{I}_1} \log \hat{p}_{\mathbf{X}|H,\mathbf{w}}(\mathbf{x}^{(k)}|1, \mathbf{w}) \\ &= -\sum_{k \in \mathcal{I}_0} \left(\frac{\log(2\pi v)}{2} + \frac{(x - m_0)^2}{2v} \right) - \sum_{k \in \mathcal{I}_1} \left(\frac{\log(2\pi v)}{2} + \frac{(x - m_1)^2}{2v} \right) \\ &= -\left(\frac{K}{2} \log(2\pi v) - \sum_{k \in \mathcal{I}_0} \frac{(x - m_0)^2}{2v} - \sum_{k \in \mathcal{I}_1} \frac{(x - m_1)^2}{2v} \right) \end{aligned} \tag{4.14}$$

Derivando respecto a cada uno de los parámetros, igualando a cero y despejando, resultan los estimadores

$$m_0^* = \frac{1}{K_0} \sum_{k \in \mathcal{I}_0} x_k \tag{4.15}$$

$$m_1^* = \frac{1}{K_1} \sum_{k \in \mathcal{I}_1} x_k \tag{4.16}$$

$$v^* = \frac{1}{K} \left(\sum_{k \in \mathcal{I}_0} (x^{(k)} - m_0^*)^2 + \sum_{k \in \mathcal{I}_1} (x^{(k)} - m_1^*)^2 \right) \tag{4.17}$$

Observe que, en el ejemplo, las verosimilitudes de las hipótesis tienen un parámetro común, v, y otro parámetro específico de cada distribución (m_0 y m_1). Como consecuencia, v^* depende de todas las muestras de entrenamiento, mientras que m_0 y m_1 solamente dependen de las muestras de una categoría.

En general, si todos los parámetros son específicos de una de las verosimilitudes, los estimadores se pueden calcular separadamente por cada categoría, de tal modo que, si llamamos $\mathbf{w}_h$ al vector que contiene todos los parámetros de la verosimilitud de la hipótesis h, puede reemplazarse (4.5) por

$$\mathbf{w}_h^* = \arg \max_{\mathbf{w}_h} \sum_{k \in \mathcal{I}_h} \log \hat{p}_{\mathbf{X}|H,\mathbf{w}_h}(\mathbf{x}^{(k)}|h, \mathbf{w}_h) \tag{4.18}$$

donde $\mathcal{I}_h$ es, como en el ejemplo, el conjunto de índices de las muestras de la categoría h (es decir, el conjunto de valores de k para los que $h^{(k)} = h$).

5

Filtrado Lineal

5.1 Introducción

Un problema común en estimación es el de querer determinar los coeficientes de un filtro lineal con M coeficientes a partir de la sola observación de las entradas y salidas de este. A esta tarea, así como a otras relacionadas, se la conoce con el nombre genérico de "filtrado lineal". En este bloque mostraremos como las técnicas descritas en el bloque B1 pueden ser usadas para diseñar estimadores ML, MAP, MAD y MMSE de los coeficientes de dicho filtro, así como de futuras salidas del filtro si se conocen las correspondientes entradas.

5.2 El problema de filtrado

Suponga que se utiliza un filtro de respuesta al impulso finita (FIR, finite impulse response), $s[n]$, con $s[n] = 0$, para n distino de $0, 1, \ldots, M-1$ para filtrar una señal $u[n]$. Al resultado se le suma cierto ruido gaussiano $\varepsilon[n]$ iid de media nula y varianza σ_ε^2, dando lugar a una observación $x[n]$. Es decir,

$$x[n] = u[n] * s[n] + \varepsilon[n] \tag{5.1}$$

$$= u[n]s[0] + u[n-1]s[1] + \ldots + u[n-M+1]s[M-1] + \varepsilon[n]. \tag{5.2}$$

Agrupando los coeficientes no nulos del filtro desconocido en un vector $\mathbf{s} = [s[0], s[1], \ldots, s[M-1]]^\top$ y compactando una secuencia contigua de longitud M de la señal de entrada $\mathbf{u}[n] = [u[n], u[n-1], \ldots, u[n-M+1]]^\top$, podemos decir que

$$x[n] = \mathbf{u}[n]^\top \mathbf{s} + \varepsilon[n]. \tag{5.3}$$

El problema de filtrado consiste entonces en estimar los coeficientes $\mathbf{s}$ de un filtro a partir de un conjunto de entradas y salidas observadas, así como estimar la salida x_* correspondiente a una nueva entrada $\mathbf{u}_*$.

Si disponemos de las señales $u[n]$ y $x[n]$ en el intervalo $0 \leq n \leq N-1$ y suponiendo que ambas señales son nulas para $n < 0$, dispondremos de un total de N

parejas entrada-salida, $\{\mathbf{u}[n], x[n]\}_{n=0}^{N-1}$. Podemos agrupar dichas parejas entrada-salida en las matrices $\mathbf{x}$ y $\mathbf{U}$:

$$\mathbf{x} = \begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[M-1] \\ \vdots \\ x[N-1] \end{bmatrix}_{N\times 1}, \tag{5.4}$$

$$\begin{aligned} \mathbf{U} &= [\mathbf{u}[0] \;\; \mathbf{u}[1] \;\; \ldots \;\; \mathbf{u}[M-1] \;\; \ldots \;\; \mathbf{u}[N-1]] \\ &= \begin{bmatrix} u[0] & u[1] & \ldots & u[M-1] & \ldots & u[N-1] \\ 0 & u[0] & \ldots & u[M-2] & \ldots & u[N-2] \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ 0 & 0 & \ldots & u[0] & \ldots & u[N-M] \end{bmatrix}_{M\times N}, \end{aligned} \tag{5.5}$$

lo que permitirá obtener expresiones compactas en las secciones siguientes.

Nota: A lo largo de las subsiguientes derivaciones, la señal $u[n]$ y por tanto la matriz $\mathbf{U}$ están consideradas como valores observados y deterministas, a los que están implicitamente condicionadas todas las expresiones probabilísticas.

5.3 Solución ML

El propio planteamiento del problema nos proporciona la verosimilitud de los coeficientes del filtro $\mathbf{s}$ dada la observación n-ésima:

$$p(x[n]|\mathbf{s}) = \mathcal{N}(x[n]|\mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2), \tag{5.6}$$

donde se utiliza la notación $\mathcal{N}(a|\mu, v)$ para referirnos a una fdp *normal* (Gaussiana) con v.a. a, media μ y varianza v.

Dado un conjunto de observaciones, simplemente tomamos el producto de las anteriores verosimilitudes, ya que los términos de ruido son independientes

$$p(\mathbf{x}|\mathbf{s}) = \prod_{n=M}^{N} \mathcal{N}(x[n]|\mathbf{u}[n]^\top \mathbf{s}, \sigma_\varepsilon^2) = \mathcal{N}(\mathbf{x}|\mathbf{U}^\top \mathbf{s}, \sigma_\varepsilon^2 \mathbf{I}). \tag{5.7}$$

El valor de $\mathbf{s}$ que maximiza $p(\mathbf{x}|\mathbf{s})$ es

$$\begin{aligned} \hat{\mathbf{s}}_{\text{ML}} &= \operatorname*{argmax}_{\mathbf{s}} p(\mathbf{x}|\mathbf{s}) = \operatorname*{argmax}_{\mathbf{s}} \log p(\mathbf{x}|\mathbf{s}) \\ &= \operatorname*{argmin}_{\mathbf{s}} \frac{1}{2}(\mathbf{x} - \mathbf{U}^\top \mathbf{s})^\top (\sigma_\varepsilon^2 \mathbf{I})^{-1} (\mathbf{x} - \mathbf{U}^\top \mathbf{s}) + \frac{1}{2} \log |\sigma_\varepsilon^2 \mathbf{I}| + \frac{N}{2} \log(2\pi) \\ &= \operatorname*{argmin}_{\mathbf{s}} ||\mathbf{x} - \mathbf{U}^\top \mathbf{s}||^2 \qquad (5.8) \\ &= (\mathbf{U}\mathbf{U}^\top)^{-1} \mathbf{U}\mathbf{x}. \qquad (5.9) \end{aligned}$$

El último paso es simplemente la solución least squares que se vió en el capítulo de regresión. Dicho mínimo se puede obtener fácilmente tomando el gradiente con respecto a $\mathbf{s}$, igualando a cero y despejando.

5.4 Solución Bayesiana

Para obtener un estimador Bayesiano de $\mathbf{s}$ necesitamos conocer su probabilidad a priori $p(\mathbf{s})$. Aunque en general ésta es desconocida, es sensato utilizar

$$p(\mathbf{s}) = \mathcal{N}(\mathbf{s}|\mathbf{0}, \sigma_s^2\mathbf{I}), \tag{5.10}$$

ya que considera aceptable cualquier conjunto de coeficientes reales, y supone que estos tienen media nula y una dispersión fijada por σ_s^2. También es posible fijar $\sigma_s^2 \to \infty$ para conseguir una distribución uniforme. En cualquier caso, el uso de esta distribución a priori permite obtener de manera analítica la distribución a posteriori.

Conocidas la verosimilitud $p(\mathbf{x}|\mathbf{s})$ y la distribución a priori $p(\mathbf{s})$, podemos obtener la distribución a posteriori $p(\mathbf{s}|\mathbf{x})$. Para ello, podríamos aplicar directamente el teorema de Bayes y simplificar el cociente tanto como sea posible, pero este es un proceso muy tedioso. En lugar de eso, vamos a obtener el resultado en dos pasos.

Primero vamos a encontrar la fdp conjunta de $\mathbf{s}$ y $\mathbf{x}$. Una manera sencilla de hacer esto es observar que

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \begin{bmatrix}\mathbf{I}\\ \mathbf{U}^\top\end{bmatrix}\mathbf{s} + \begin{bmatrix}\mathbf{0}\\ \boldsymbol{\varepsilon}\end{bmatrix} \text{ con } \boldsymbol{\varepsilon} = [\varepsilon[0], \ldots, \varepsilon[N-1]]^\top, \tag{5.11}$$

es decir, el vector $[\mathbf{s}^\top\ \mathbf{x}^\top]^\top$ es una combinación lineal de v.a. con fdp Gaussiana más un vector independiente de ruido blanco y Gaussiano con varianza σ_ε^2, y por tanto, conjuntamente Gaussiano. Obtener la media y varianza de $[\mathbf{s}^\top\ \mathbf{x}^\top]^\top$ es por tanto inmediato:

$$\begin{bmatrix}\mathbf{s}\\ \mathbf{x}\end{bmatrix} = \mathcal{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_s^2\mathbf{I} & \sigma_s^2\mathbf{U}\\ \sigma_s^2\mathbf{U}^\top & \sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I}\end{bmatrix}\right) \tag{5.12}$$

y utilizando la fórmula de condicionamiento en Gaussianas vista en los apuntes de estimación (B1), tenemos que

$$p(\mathbf{s}|\mathbf{x}) = \mathcal{N}(\mathbf{s} \mid \sigma_s^2\mathbf{U}(\sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{x},\ \ \sigma_s^2\mathbf{I} - \sigma_s^2\mathbf{U}(\sigma_s^2\mathbf{U}^\top\mathbf{U} + \sigma_\varepsilon^2\mathbf{I})^{-1}\mathbf{U}\sigma_s^2), \tag{5.13}$$

lo cual, usando el lema de inversión de matrices y algo de algebra, puede operarse hasta obtener la siguiente expresión, más simple y computacionalmente más eficiente cuando $M < N$:

$$p(\mathbf{s}|\mathbf{x}) = \mathcal{N}(\mathbf{s} \mid \mathbf{P}\mathbf{U}\mathbf{x},\ \ \sigma_\varepsilon^2\mathbf{P}), \tag{5.14}$$

donde hemos definido $\mathbf{P} = (\mathbf{U}\mathbf{U}^\top + \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}$. Esto nos proporciona los siguientes estimadores para $\mathbf{s}$:

$$\hat{\mathbf{s}}_{\text{MMSE}} = \hat{\mathbf{s}}_{\text{MAP}} = \hat{\mathbf{s}}_{\text{MAD}} = \mathbf{P}\mathbf{U}\mathbf{x} \tag{5.15}$$

Nótese que suponer una distribución a priori uniforme (usando $\sigma_s^2 \to \infty$ en la expresión anterior) convierte la solución MAP en la solución ML obtenida anteriormente.

5.4.1 Predicción probabilística de la salida del filtro

Una vez resuelto obtenidos diversos estimadores del filtro $\mathbf{s}$, pasamos ahora a plantearnos el problema de estimar una nueva salida x_* correspondiente a una nueva entrada $\mathbf{u}_*$. Siguiendo con la perspectiva Bayesiana, obtendremos la fdp a posteriori de la variable a estimar, x_*, a la vista de las salidas ya observadas, $\mathbf{x}$. Es decir, queremos calcular $p(x_*|\mathbf{x})$.

En primer lugar, hay que notar que $\mathbf{x}$, x_* y $\mathbf{s}$ son conjuntamente Gaussianas. Esto se sigue de la ecuación (5.12), que puede ampliarse a cualquier número arbitrario de salidas, incluyendo x_*. Esto implica necesariamente que $\mathbf{x}$ y x_* son conjuntamente Gaussianas (al marginalizar $\mathbf{s}$) y finalmente que $p(x_*|\mathbf{x})$ debe ser Gaussiana. Dado que

$$x_* = \mathbf{u}_*^\top \mathbf{s} + \varepsilon_* \tag{5.16}$$

es una transformación lineal de $\mathbf{s}$ más ruido blanco independiente, podemos facilmente calcular la media $\mathbb{E}[x_*|\mathbf{x}]$ y varianza $\mathbb{V}[x_*|\mathbf{x}]$ de dicha Gaussiana usando $p(\mathbf{s}|\mathbf{x})$, que ya conocemos, dando lugar a

$$p(x_*|\mathbf{x}) = \mathcal{N}(x_* \mid \mathbf{u}_*^\top \mathbf{P}\mathbf{U}\mathbf{x},\ \ \sigma_\varepsilon^2 + \sigma_\varepsilon^2 \mathbf{u}_*^\top \mathbf{P}\mathbf{u}_*). \tag{5.17}$$

que inmediatamente nos proporciona los siguientes estimadores para x_*:

$$\hat{x}_{*\text{MMSE}} = \hat{x}_{*\text{MAP}} = \hat{x}_{*\text{MAD}} = \mathbf{u}_*^\top \mathbf{P}\mathbf{U}\mathbf{x} = \mathbf{u}_*^\top \hat{\mathbf{s}}_{\text{MMSE}}. \tag{5.18}$$

Se observa por tanto, que para obtener los diversos estimadores mencionados de la nueva salida x_* es suficiente con conocer la nueva entrada $\mathbf{u}_*$ y el estimador $\mathbf{s}_{\text{MMSE}}$.

5.5 Cálculo online

Es posible obtener las soluciones anteriores de manera online, es decir, a medida que se obtienen nuevas parejas entrada-salida. Si bien se podrían repetir los cálculos completos cada vez que llega una nueva muestra, a menudo existen maneras más eficientes de hacer esto.

Observe que la estimación de $\mathbf{s}$ mediante las ecs. (5.9) o (5.15) requiere invertir una matriz de tamaño $M \times M$. Esto tiene un coste $\mathcal{O}(M^3)$, es decir, si doblamos el tamaño del filtro M, multiplicamos por ocho su coste computacional. Supongamos ahora que se desea estimar $\mathbf{s}$ a medida que se reciben nuevas parejas entrada-salida, es decir, se nos da primero $\{u[0], x[0]\}$, luego $\{u[1], x[1]\}$ y así sucesivamente. En este caso, podríamos reutilizar los resultados de la estimación anterior para calcular la nueva estimación actualizada de $\mathbf{s}$, reduciendo así el coste $\mathcal{O}(M^3)$ que tendría un método "ingenuo" que simplemente recalcule todo de nuevo cada vez que llega una muestra.

5.5.1 Solución Bayesiana

Se puede obtener de manera exacta $\hat{\mathbf{s}}_{\text{MMSE}}$ a medida que se dispone de más muestras (N aumenta) sin necesidad de rehacer todos los calculos, reusando la solución anterior. Para ello, se define $\mathbf{P}^{(N)} = (\mathbf{U}\mathbf{U}^\top + \frac{\sigma_\varepsilon^2}{\sigma_s^2}\mathbf{I})^{-1}$, $\mathbf{r}^{(N)} = \mathbf{U}\mathbf{x}$ y se usa el siguiente cálculo recursivo (la primera ecuación corresponde a la aplicación directa del lema de inversión de matrices a la actualización de $\mathbf{P}$):

$$\begin{aligned}\mathbf{P}^{(N+1)} &= \mathbf{P}^{(N)} - \frac{\mathbf{P}^{(N)}\mathbf{u}[N+1]\mathbf{u}[N+1]^\top\mathbf{P}^{(N)}}{1+\mathbf{u}[N+1]^\top\mathbf{P}^{(N)}\mathbf{u}[N+1]}\\ \mathbf{r}^{(N+1)} &= \mathbf{r}^{(N)} + \mathbf{u}[N+1]x[N+1]\\ \mathbf{s}^{(N+1)} &= \mathbf{P}^{(N+1)}\mathbf{r}^{(N+1)},\end{aligned}$$

que sólo tiene un coste $\mathcal{O}(M^2)$ por paso (a diferencia de aplicar la ecuación original completa en cada paso, que tendría coste $\mathcal{O}(M^3)$). A este algoritmo se le llama recursive least squares (RLS).

5.5.2 Solución ML

Se puede obtener una aproximación a $\hat{\mathbf{s}}_{\text{ML}}$ online con coste computacional $\mathcal{O}(M)$ sin más que notar que

$$\hat{\mathbf{s}}_{\text{ML}} = \underset{\mathbf{s}}{\operatorname{argmax}}\, p(\mathbf{x}|\mathbf{s}) = \underset{\mathbf{s}}{\operatorname{argmin}}\, ||\mathbf{x} - \mathbf{U}^\top\mathbf{s}||^2 \tag{5.19}$$

y a continuación usar gradiente estocástico para minimizar $||\mathbf{x} - \mathbf{U}^\top\mathbf{s}||^2$.

Nótese que

$$||\mathbf{x} - \mathbf{U}^\top\mathbf{s}||^2 = \sum_{n=0}^{N-1}(x[n] - \mathbf{u}[n]^\top\mathbf{s})^2, \tag{5.20}$$

por lo que un método de descenso por gradiente calcularía el gradiente de dicha expresión y desplazaría iterativamente la estimación del mínimo en la dirección opuesta al gradiente en cada paso. Un descenso por gradiente estocástico realiza la misma operación, pero considerando únicamente uno de los sumandos del mencionado sumatorio en cada paso. Así, la actualización de coeficientes que debe iterarse para efectuar la minimización es en este caso

$$\hat{\mathbf{s}}^{(N+1)} = \hat{\mathbf{s}}^{(N)} + \mu\left(x[n] - \mathbf{u}[n]^\top\hat{\mathbf{s}}^{(N)}\right)\mathbf{u}[n], \tag{5.21}$$

donde μ es un paso de adaptación "suficientemente pequeño". A este algoritmo se le llama least mean squares (LMS).

5.6 Filtro de Wiener

El filtro de Wiener $\mathbf{s}_{\text{Wiener}}$ es el filtro que minimiza el error cuadrático esperado entre una salida deseada $x[n]$ y la salida producida al ser utilizado para filtrar la entrada

$u[n]$. En este apartado, tanto $x[n]$ como $u[n]$ se consideran señales de media nula y $u[n]$ se trata como un proceso estocástico y no como una señal determinista, como se ha venido haciendo hasta ahora.

Este problema se puede plantear como un problema de estimación lineal de mínimo error cuadrático medio (MMSE), por lo que se pueden usar la formlación de la sección B1 para dar lugar a la siguiente solución:

$$\mathbf{s}_{\text{Wiener}} = \mathbf{R}_{uu}^{-1}\mathbf{r}_{ux}, \tag{5.22}$$

donde $\mathbf{R}_{uu}$ es la matriz de autocorrelación de la señal de entrada $u[n]$ y $\mathbf{r}_{ux}$ es el vector de correlación cruzada entre $u[n]$ y $x[n]$. Desafortunadamente, estas dos cantidades son desconocidas en general, por lo que en la mayoría de las ocasiones, el filtro de Wiener no puede calcularse. Sin embargo, es frecuente usar la expresión anterior usando estimaciones muestrales para la matriz de correlación $\hat{\mathbf{R}}_{uu} = \frac{1}{N}\mathbf{U}\mathbf{U}^\top$ y el vector de correlación cruzada $\hat{\mathbf{r}}_{ux} = \frac{1}{N}\mathbf{U}\mathbf{x}$. El resultado es una aproximación al filtro de Wiener $\hat{\mathbf{s}}_{\text{Wiener}} = \hat{\mathbf{R}}_{uu}^{-1}\hat{\mathbf{r}}_{ux}$ que minimiza el error cuadrático muestral (a menudo llamada "estimación least-squares") y que coincide con la solución ML, es decir $\hat{\mathbf{s}}_{\text{Wiener}} = \hat{\mathbf{s}}_{\text{ML}}$.

A medida que el número de muestras disponibles para la estimación de los estadísticos $\mathbf{R}_{uu}$ y $\mathbf{r}_{ux}$ aumenta, dichas estimaciones se vuelven más precisas, de manera que $\hat{\mathbf{s}}_{\text{Wiener}}$ y por tanto $\hat{\mathbf{s}}_{\text{ML}}$ coinciden asintóticamente con el verdadero filtro de Wiener.

5.7 Problemas

5.1. Considere la siguiente secuencia

$$u[1]\ldots u[7] \equiv 0.7,\ -0.1,\ 0.7,\ -0.2,\ -0.1,\ 1.5,\ -1.1$$

que se alimenta como entrada a un filtro lineal de tres coeficientes $\mathbf{s} = [s_1, s_2, s_3]^\top$. Se conocen los siguientes elementos de la secuencia de salida, (corrompidos con ruido Gaussiano de varianza 0.25):

$$x[1]\ldots x[6] \equiv -0.60,\ 1.13,\ 0.57,\ 0.42,\ 1.25,\ -2.58$$

a) ¿Cuál es la estimación ML de $\mathbf{s}$? (filtro de Wiener basado en estadísticos aproximados).
b) Utilice el filtro obtenido para predecir $x[7]$, $\hat{x}_{\text{ML}}$.
c) Calcule las estimación MMSE, MAP y MAD de $\mathbf{s}$ asumiendo que la pdf a priori de sus componentes es $s_i \sim \mathcal{N}(0, 1)$.
d) Obtenga la estimación MMSE de $x[7]$, $\hat{x}_{\text{MMSE}}$.
e) Calcule el error cuadrático esperado en la predicción b). (Es decir, la esperanza de $(\hat{x}_{\text{ML}} - x[7])^2$ a la vista de los datos disponibles).
f) Calcule el error cuadrático esperado en la predicción d). (Es decir, la esperanza de $(\hat{x}_{\text{MMSE}} - x[6])^2$ a la vista de los datos disponibles).

6

Soluciones de los problemas

6.1 Problemas del Capítulo 1

1.1 Los estimadores buscados son:

$$\hat{s}_{\text{MMSE}} = 1/x^2$$

$$\hat{s}_{\text{MAD}} = \frac{\ln 2}{x^2}$$

$$\hat{s}_{\text{MAP}} = 0$$

1.2 Los estimadores buscados son:

$$\hat{s}_{\text{MMSE}} = 1/x$$

$$\hat{s}_{\text{MAD}} = \frac{\ln 2}{x}$$

$$\hat{s}_{\text{MAP}} = 0$$

La Figura 6.1 representa la distribución a posteriori de S para $X = 2$, e indica el valor de los diferentes estimadores estudiados para dicha observación.

1.3

a) $\hat{S}_{\text{LMSE}} = 0.5X - 0.5$.
b) $\mathbb{E}\left\{\left(S - \hat{S}_{\text{LMSE}}\right)^2\right\} = 0.75$.

1.4

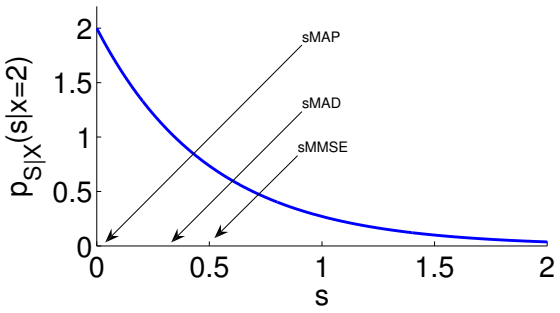


Fig. 6.1. Estimación de una variable aleatoria S cuya distribución dado $X = 2$ es exponencial.

a) $\hat{S}_{\text{MMSE}} = \dfrac{X}{2}$.

b) $\mathbb{E}\left\{\left(S - \hat{S}_{\text{MMSE}}\right)^2\right\} = 0.$

1.5

$\hat{r}_{ML} = -\frac{1}{64}\sum_{k=1}^{64} \ln y^{(k)}$

1.6

a) $\hat{\alpha}_{ML} = 0.56,$
$\hat{v}_{ML} = 0.029$

b) $A_{min} = 1.91.$

6.2 Problemas del Capítulo 3

3.1

a)

$$\hat{d}_{\text{ML}} = \begin{cases} 0, & \text{si } x_1 < 0.5 \quad \text{y} \quad x_1 + x_2 < 1 \\ 1, & \text{si } x_1 > 0.5 \quad \text{y} \quad x_1 > x_2 \\ 2, & \text{resto} \end{cases}$$

b) Las regiones de decisión se muestran en la figura:

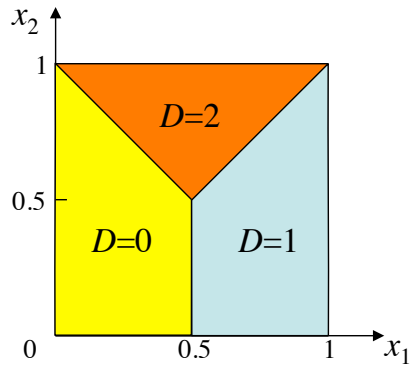


Fig. 6.2. Regiones de decisión.

3.2

a)

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \frac{1}{2}$$

b)

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \frac{5}{6}$$

c)

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \frac{2}{3}$$

d) $P_{\text{FA}} = (1-\eta)^2$, $P_{\text{M}} = \eta$, $P_e = \frac{3}{4}(1-\eta)^2 + \frac{1}{4}\eta$

e)

	η	P_{M}	P_{FA}	P_e
ML	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{5}{16}$
MAP	$\frac{5}{6}$	$\frac{5}{6}$	$\frac{1}{36}$	$\frac{11}{48}$
Bayesiano	$\frac{2}{3}$	$\frac{2}{3}$	$\frac{1}{9}$	$\frac{1}{4}$

f)

	r_ϕ
ML	$\frac{7}{16}$
MAP	$\frac{21}{48}$
Bayesiano	$\frac{5}{12}$

3.3

a)

$$D = \begin{cases} 1, & \text{si } \dfrac{1}{3} < |x| < 1 \\ 0, & \text{resto.} \end{cases}$$

b) $P_{\text{FA}} = \dfrac{2}{9}$, $P_{\text{M}} = \dfrac{1}{27}$, $P_e = \dfrac{7}{54}$

3.4

a) El decisor ML es

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \ln \frac{2}{\eta}$$

b) $P_D = P_{\text{FA}}(2 - P_{\text{FA}})$

3.5

a) El decisor LRT es

$$x \underset{D=0}{\overset{D=1}{\gtrless}} \frac{\eta}{2}$$

b) $P_D = P_{\text{FA}}(2 - P_{\text{FA}})$. El decisor ML opera en el punto $(P_{\text{FA}}, P_{\text{D}}) = (0.5, 0.75)$. El decisor NP opera en $(P_{\text{FA}}, P_{\text{D}}) = (0.1, 0.19)$

3.6 Se trata de un problema de decisión con costes: $c_{11} = 1$, $c_{10} = 1$, $c_{01} = 81$ y $c_{00} = -3$ y probabilidades a priori $P_H(0) = 100/101$, $P_H(1) = 1/101$,

a) El decisor óptimo es

$$x_1 \underset{D=0}{\overset{D=1}{\gtrless}} \eta = 2 \ln 10 \approx 4.61$$

b) El coste medio por unidad de producto será

$$\overline{C} = -\frac{223}{101} \approx -2.21$$

luego el beneficio diario esperado es

$$B = \frac{2230000}{101} \approx 22100$$

c) El coste medio por unidad de producto es

$$C' = -\frac{1}{101}\left(218.5 + 90 \cdot 40^{-\frac{1}{3}}\right) \approx -2.42$$

y el tiempo medio que tardaría E en amortizar la máquina será

$$T = \frac{6000}{\overline{C} - \overline{C}'} \approx 28600 \text{ días}$$

3.7 Se obtiene un ejemplo sencillo buscando un problema de decisión bidimensional ($\mathbf{x} = (x_1, x_2)$) que conduzca a una frontera de decisión de la forma $x_2 = ax_1^2 + b$ (siendo a y b constantes arbitrarias). Por ejemplo, para $\mathbf{m}_0 = \mathbf{0}$, $\mathbf{m}_1 = (0,1)^{\mathrm{T}}$, y matrices de varianzas

$$\mathbf{V}_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

y

$$\mathbf{V}_1 = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix}$$

3.8

a) El criterio ML indica que se transmitió el símbolo s_1 ($0.6 > \frac{1}{2}$). $P_{\text{FA}} = 0.3085$ y $P_{\text{M}} = 0.3085$.
b) Como $0.6 < 1.193$, el decisor MAP decide s_0 con $P_{\text{FA}} = 0.1164$ y $P_{\text{M}} = 0.5763$.
c) Como $0.6 < 0.674$, el test de Neyman-Parson decide que se transmitió s_0 con una probabilidad de pérdida de 0.3722.
d)

$$\frac{1}{K}\sum_{k=1}^{K} x^{(k)} \underset{D=0}{\overset{D=1}{\gtrless}} \frac{1}{2}$$

e) $K = 5$: $P_{\text{FA}} = P_{\text{M}} = 0.1318$. $K = 10$: $P_{\text{FA}} = P_{\text{M}} = 0.057$.

6.3 Problemas del Capítulo 5

5.1 Primero, escribimos los datos en la forma matricial vista en la teoría:

$$\mathbf{U} = \begin{bmatrix} 0.7 & -0.2 & -0.1 & 1.5 \\ -0.1 & 0.7 & -0.2 & -0.1 \\ 0.7 & -0.1 & 0.7 & -0.2 \end{bmatrix}; \quad \mathbf{x} = \begin{bmatrix} 0.57 \\ 0.42 \\ 1.25 \\ -2.58 \end{bmatrix}; \quad \mathbf{u}_* = [-1.1,\ 1.5,\ -0.1]^\top,$$

donde $\mathbf{u}_*$ es el vector de entrada que daría lugar, al ser multiplicado por el vector de pesos del filtro y corrompido con ruido Gaussiano de varianza 0.25, a $x[7]$.

a)

$$\hat{\mathbf{s}}_{\text{ML}} = (\mathbf{U}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{x} = \{\mathbf{U}^\top \backslash \mathbf{x} \text{ en MatLAB/Octave}\} = [-1.35,\ 0.57,\ 2.02]^\top$$

b)

$$\hat{x}_{\text{ML}} = \mathbf{u}_*^\top \hat{\mathbf{s}}_{\text{ML}} = [-1.1,\ 1.5,\ -0.1] \cdot [-1.35,\ 0.57,\ 2.02]^\top = 2.14$$

c)

$$\hat{\mathbf{s}}_{\text{MMSE}} = \hat{\mathbf{s}}_{\text{MAP}} = \hat{\mathbf{s}}_{\text{MAD}} = (\mathbf{U}\mathbf{U}^\top + 0.25\mathbf{I})^{-1}\mathbf{U}\mathbf{x} = [-1.25,\ 0.28,\ 1.56]^\top$$

d)

$$\hat{x}_{\text{MMSE}} = \mathbf{u}_*^\top \hat{\mathbf{s}}_{\text{MMSE}} = [-1.1,\ 1.5,\ -0.1] \cdot [-1.25,\ 0.28,\ 1.56]^\top = 1.64$$

e)

$$\begin{aligned} &\mathrm{e}[(\hat{x}_{\text{ML}} - x[7])^2 | \mathbf{U}, \mathbf{x}, \mathbf{u}_*] = \\ &= \int (\hat{x}_{\text{ML}} - x[7])^2 \mathcal{N}(x[7] |\ \mathbf{u}_*^\top \hat{\mathbf{s}}_{\text{MMSE}},\ 0.25 + 0.25\mathbf{u}_*^\top (\mathbf{U}\mathbf{U}^\top + 0.25\mathbf{I})^{-1}\mathbf{u}_*) dx[7] \\ &= \int (\hat{x}_{\text{ML}} - x[7])^2 \mathcal{N}(x[7] \mid 1.64,\ 1.004) dx[7] = (2.14 - 1.64)^2 + 1.004 = 1.254 \end{aligned}$$

f)

$$\begin{aligned} &\mathrm{e}[(\hat{x}_{\text{MMSE}} - x[7])^2 | \mathbf{U}, \mathbf{x}, \mathbf{u}_*] = \\ &= \int (\hat{x}_{\text{MMSE}} - x[7])^2 \mathcal{N}(x[7] |\ \mathbf{u}_*^\top \hat{\mathbf{s}}_{\text{MMSE}},\ 0.25 + 0.25\mathbf{u}_*^\top (\mathbf{U}\mathbf{U} + 0.25\mathbf{I})^{-1}\mathbf{u}_*) dx[7] \\ &= \int (\hat{x}_{\text{MMSE}} - x[7])^2 \mathcal{N}(x[7] \mid 1.64,\ 1.004) dx[7] = (1.64 - 1.64)^2 + 1.004 = 1.004 \end{aligned}$$

References

1. Hayes M H (1996) Statistical Digital Signal Processing and Modeling. John Wiley and Sons, New York, EE.UU.
2. Oppenheim A, Schaffer R (1999) Discrete-Time Signal Processing 2nd Ed. Prentice Hall, New York, EE.UU.
   Thesis, Columbia University, New York