

APÉNDICE I

I. BREVES NOCIONES DE PROBABILIDAD Y ESTIMAS DE MÁXIMA VEROSIMILITUD

I.1 PROBABILIDAD TOTAL Y LA REGLA DE BAYES

Consideremos M eventos A_i , $i = 1, 2, \dots, M$ de modo que la suma de sus probabilidades sea uno, es decir $\sum_{i=1}^M P(A_i) = 1$, entonces la probabilidad de un evento arbitrario B viene dada por

$$P(B) = \sum_{i=1}^M P(B / A_i) P(A_i) \quad (\text{I.1})$$

donde $P(B / A_i)$ expresa la probabilidad condicional de B asumiendo A_i , que se define como,

$$P(B / A) = \frac{P(B, A)}{P(A)} \quad (\text{I.2})$$

y $P(B, A)$ es la probabilidad conjunta de dos eventos. La ecuación (I.1) se conoce como el *teorema de probabilidad total*. A partir de la definición en (I.2) obtenemos la *regla de Bayes*

$$P(B / A) P(A) = P(A / B) P(B) \quad (I.3)$$

La regla de Bayes se extiende fácilmente a variables aleatorias o vectores descritos por funciones de densidad de probabilidad y tenemos,

$$p(\mathbf{x} / A) P(A) = P(A / \mathbf{x}) p(\mathbf{x}) \quad (I.4)$$

y

$$p(\mathbf{x} / \mathbf{y}) p(\mathbf{y}) = p(\mathbf{y} / \mathbf{x}) p(\mathbf{x}) \quad (I.5)$$

y finalmente

$$p(\mathbf{x}) = \sum_{i=1}^M p(\mathbf{x} / A_i) P(A_i) \quad (I.6)$$

I.2 MEDIA Y VARIANZA

Sea $p(x)$ la función de densidad de probabilidad describiendo la variable aleatoria x . Su media y su varianza están definidas como,

$$E[x] = \int_{-\infty}^{+\infty} xp(x)dx, \quad \sigma_x^2 = \int_{-\infty}^{+\infty} (x - E[x])^2 p(x)dx \quad (I.7)$$

I.3 INDEPENDENCIA ESTADÍSTICA

Dos (o más) variables aleatorias x e y son estadísticamente independientes si y sólo si

$$p(x, y) = p(x)p(y) \quad (I.8)$$

en este caso $E[xy] = E[x]E[y]$, esto se puede generalizar a más de dos variables.

I.4 FUNCIÓN DE DENSIDAD DE PROBABILIDAD MULTIVARIABLE GAUSSIANA O NORMAL

Se trata de una generalización de la función de densidad de probabilidad de una sola variable,

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |C|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \mathbf{m})^t C^{-1} (\mathbf{x} - \mathbf{m}) \right\} \quad (\text{I.9})$$

donde $\mathbf{x} = (x_1, x_2, \dots, x_d)^t$ y \mathbf{m} el vector media $E[(x_1, x_2, \dots, x_d)^t] = (m_1, m_2, \dots, m_d)^t$ y C la matriz de covarianza. $|C|$ es el determinante de dicha matriz de covarianza.

$$C = E[(\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^t] \quad (\text{I.10})$$

se dice que \mathbf{x} es una distribución normal y se expresa $N(\mathbf{m}, C)$.

El término $d_M^2(\mathbf{x}, \mathbf{m}) = (\mathbf{x} - \mathbf{m})^t C^{-1} (\mathbf{x} - \mathbf{m})$ en (I.9) es la denominada distancia de Mahalanobis al cuadrado de \mathbf{x} a \mathbf{m} y cuando la matriz de covarianza es la identidad, la distancia de Mahalanobis al cuadrado resulta ser la distancia Euclídea al cuadrado.

Para el caso unidimensional $d=1$, la matriz de covarianza es la varianza σ^2 y la función de densidad Gaussiana toma la forma,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(x - m)^2}{2\sigma^2} \right\} \quad (\text{I.11})$$

La figura I.1 muestra el gráfico de dos Gaussianas para la misma media y diferentes varianzas

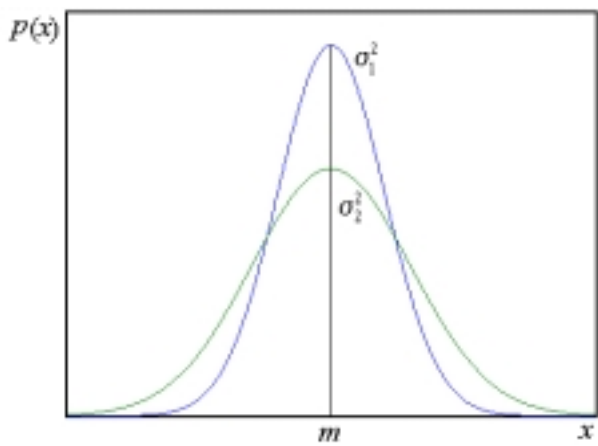


Figura I.1 Dos Gaussianas con la misma media m y diferentes varianzas,

$$\text{con } \sigma_1^2 < \sigma_2^2$$

Para el caso d -dimensional general, la matriz de covarianza tiene la forma,

$$C = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2d} \\ \vdots & \vdots & \cdots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{pmatrix} \quad (\text{I.12})$$

donde $\sigma_i^2 = E[(x_i - m_i)^2]$, $\sigma_{ij} = \sigma_{ji} = E[(x_i - m_i)(x_j - m_j)]$. Por tanto, la diagonal principal consta de las respectivas varianzas de los elementos del vector aleatorio y los elementos fuera de dicha diagonal son las respectivas covarianzas entre los elementos del vector aleatorio. Obsérvese, que si las variables aleatorias x_i son estadísticamente independientes, la media del producto es igual al producto de las medias, esto es,

$$E[(x_i - m_i)(x_j - m_j)] = E[(x_i - m_i)]E[(x_j - m_j)] = 0 \quad (\text{I.13})$$

en este caso la matriz de covarianza resulta ser diagonal. No obstante, una matriz de covarianza diagonal, en general no significa que las variables sean estadísticamente independientes. En el caso de densidades Gaussianas multivariable lo opuesto es válido también. En efecto, si la matriz de covarianza es diagonal entonces es sencillo ver que

$$p(\mathbf{x}) = \prod_{i=1}^d p_i(x_i) \quad (\text{I.14})$$

donde

$$p_i(x_i) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x_i - m_i)^2}{2\sigma_i^2}\right\} \quad (\text{I.15})$$

que es la Gaussiana univariable describiendo la i -ésima variable. Entonces la densidad de probabilidad conjunta es el producto de las individuales, que es la definición de independencia estadística.

Aun cabe hacer algunas observaciones relativas a la matriz de covarianza por su utilidad en reconocimiento de patrones. Dado un vector de atributos

$\mathbf{x} = (x_1, x_2, \dots, x_d)^t$ el coeficiente de correlación entre dos elementos cualesquiera x_i, x_j de \mathbf{x} , viene dado por la siguiente expresión,

$$r_{ij} = \frac{c_{ij}}{\sigma_i \sigma_j} \quad (\text{I.16})$$

Siendo c_{ij} un elemento genérico de la matriz de covarianza definida en (I.12) y σ_i, σ_j las desviaciones típicas de las características x_i, x_j , respectivamente. Por consiguiente, se puede expresar la matriz de covarianza de la siguiente forma,

$$C = \begin{pmatrix} \sigma_1^2 & r_{12}\sigma_1\sigma_2 & \cdots & r_{1n}\sigma_1\sigma_n \\ r_{12}\sigma_1\sigma_2 & \sigma_2^2 & \cdots & r_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \cdots & \vdots \\ r_{1n}\sigma_1\sigma_n & r_{2n}\sigma_2\sigma_n & \cdots & \sigma_n^2 \end{pmatrix} \quad (\text{I.17})$$

Si el coeficiente de correlación r_{ij} es cero, se dice que las características x_i, x_j están incorreladas. Esto es una propiedad importante, ya que interesa escoger los atributos de tal forma que la matriz de covarianza sea diagonal pura para lograr que el coeficiente de correlación entre atributos sea nulo y así conseguir que todos los atributos tengan un alto poder discriminante.

I.5 ESTIMAS DE MÁXIMA VEROSIMILITUD PARA DENSIDADES DE PROBABILIDAD MIXTAS

Esta sección debe estudiarse en conjunción con el Capítulo 13, ya que muchos de los conceptos utilizados aquí son definidos en dicho capítulo, por ello haremos referencia al mismo. En cualquier caso, una detallada descripción de este procedimiento lo encontramos en Duda y Hart (1973).

Vamos a considerar un procedimiento *no supervisado* que utiliza muestras no etiquetadas, es decir veremos qué se puede hacer cuando lo que tenemos es un conjunto de muestras sin conocer su clasificación.

Comenzamos suponiendo que conocemos la estructura completa de probabilidad para el problema, con la única excepción de los valores de algunos parámetros. Para ser más precisos, hacemos las siguientes suposiciones:

- 1) Se dispone de un conjunto $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ de n muestras
- 2) Las muestras provienen de un conjunto de clases conocidas, c_j

- 3) Las probabilidades a priori $P(y=c_j)$ de pertenencia a una determinada clase se conocen, $j = 1, \dots, c$
- 4) Las formas de las funciones de densidad de probabilidad condicional $p(\mathbf{x} / c_j, \mathbf{w}_j)$ son conocidas, $j=1..N$
- 5) Lo único desconocido son los valores de los parámetros \mathbf{w}_j para los c vectores.

Los patrones de estímulo se supone que se obtienen seleccionando una clase c_j con probabilidad $P(y=c_j)$ y luego seleccionando una \mathbf{x} de acuerdo a la ley de probabilidad $p(\mathbf{x} / c_j, \mathbf{w}_j)$. Por tanto, la función de densidad de probabilidad para las muestras viene dada por

$$p(\mathbf{x} / \mathbf{w}) = \sum_{j=1}^c p(\mathbf{x} / c_j, \mathbf{w}_j) P(y = c_j) \quad (\text{I.18})$$

donde $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_c)$. Una función de densidad de esta forma se denomina *densidad mixta*. Las probabilidades de densidad de probabilidad condicional $p(\mathbf{x}/c_j, \mathbf{w}_j)$ son las densidades componentes y las probabilidades a priori $P(y=c_j)$ son los *parámetros de la mezcla*. Estos parámetros de la mezcla se pueden incluir entre los parámetros desconocidos.

Utilizando *máxima verosimilitud* supongamos que tenemos como siempre un conjunto $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ de n muestras no etiquetadas y consideremos la densidad mixta (I.18). A partir de (13.3), la verosimilitud de las muestras observadas es por definición

$$P(X / \mathbf{w}) = \prod_{k=1}^n p(\mathbf{x}_k, \mathbf{w}) \quad (\text{I.19})$$

La estima de máxima verosimilitud \mathbf{w}^* es el valor de \mathbf{w} que maximiza $P(X / \mathbf{w})$. Teniendo en cuenta los conceptos del capítulo 13 y suponiendo que $P(X / \mathbf{w})$ es una función diferenciable de \mathbf{w} ,

$$l = \sum_{k=1}^n \ln p(\mathbf{x}_k / \mathbf{w}) \quad (\text{I.20})$$

y

$$\nabla_{\mathbf{w}_i} l = \sum_{k=1}^n \frac{1}{p(\mathbf{x}_k / \mathbf{w})} \nabla_{\mathbf{w}_i} \left[\sum_{j=1}^c p(\mathbf{x}_k / c_j, \mathbf{w}_j) P(c_j) \right] \quad (\text{I.21})$$

donde l representa el logaritmo de la verosimilitud y $\nabla_{\mathbf{w}_i} l$ el gradiente de l con respecto a \mathbf{w}_i ; por simplicidad notacional hemos sustituido la expresión $P(y=c_j)$ por $P(c_j)$.

Si suponemos que los elementos de \mathbf{w}_i y \mathbf{w}_j son funcionalmente independientes si $i \neq j$, y si introducimos la probabilidad a posteriori,

$$P(c_i / \mathbf{x}_k, \mathbf{w}) = \frac{p(\mathbf{x}_k / c_i, \mathbf{w}_i) P(c_i)}{p(\mathbf{x}_k / \mathbf{w})} \quad (\text{I.22})$$

con esta expresión, el gradiente del logaritmo de la verosimilitud se puede escribir de la siguiente forma interesante

$$\nabla_{\mathbf{w}_i} l = \sum_{k=1}^n P(c_i / \mathbf{x}_k, \mathbf{w}) \nabla_{\mathbf{w}_i} \ln p(\mathbf{x}_k / c_i, \mathbf{w}_i) \quad (\text{I.23})$$

puesto que el gradiente debe ser cero en los \mathbf{w}_i que maximizan l , la estima de máxima verosimilitud \mathbf{w}_i^* debe satisfacer las condiciones

$$\sum_{k=1}^n P(c_i / \mathbf{x}_k, \mathbf{w}_i^*) \nabla_{\mathbf{w}_i} \ln p(\mathbf{x}_k / c_i, \mathbf{w}_i^*) = 0, \quad i = 1, \dots, c \quad (\text{I.24})$$

De forma inversa, entre las soluciones a esas ecuaciones para \mathbf{w}_i^* encontramos la solución de máxima verosimilitud.

Los resultados obtenidos se pueden generalizar para incluir las probabilidades a priori $P(c_j)$ como parámetros desconocidos. En este caso, la búsqueda para el máximo valor de $P(X / \mathbf{w})$ se extiende sobre \mathbf{w} y $P(c_j)$ con las restricciones

$$P(c_i) \geq 0 \quad \sum_{i=1}^c P(c_i) = 1 \quad i = 1, \dots, c \quad (\text{I.25})$$

Sean $P^*(c_i)$ y \mathbf{w}_i^* las estimas de máxima verosimilitud para $P(c_j)$ y \mathbf{w}_i . Si la función de verosimilitud es diferenciable y si $P^*(c_i) \neq 0$ para cualquier i , entonces $P^*(c_i)$ y \mathbf{w}_i^* deben satisfacer

$$P^*(c_i) = \frac{1}{n} \sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}_i^*) \quad (\text{I.26})$$

$$\sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}^*) \nabla_{\mathbf{w}_i} \ln p(\mathbf{x}_k / c_i, \mathbf{w}_i^*) = 0 \quad (\text{I.27})$$

donde

$$P^*(c_i / \mathbf{x}_k, \mathbf{w}^*) = \frac{p(\mathbf{x}_k / c_i, \mathbf{w}_i^*) P^*(c_i)}{\sum_{j=1}^c p(\mathbf{x}_k / c_j, \mathbf{w}_j^*) P^*(c_j)} \quad (\text{I.28})$$

En Duda y Hart (1973) podemos encontrar dos casos de aplicación a densidades normales, en el primer caso se consideran que los únicos parámetros desconocidos son los vectores medias y en el segundo todos son desconocidos, es decir las probabilidades a priori $P(c_j)$, los vectores medias \mathbf{m}_i y las matrices de covarianza C_i . En este último caso las estimas de máxima verosimilitud resultan ser

$$P^*(c_i) = \frac{1}{n} \sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}^*) \quad (\text{I.29})$$

$$\mathbf{m}_i^* = \frac{\sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}^*) \mathbf{x}_k}{\sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}^*)} \quad (\text{I.30})$$

$$C_i^* = \frac{\sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}^*) (\mathbf{x}_k - \mathbf{m}_i^*) (\mathbf{x}_k - \mathbf{m}_i^*)^t}{\sum_{k=1}^n P^*(c_i / \mathbf{x}_k, \mathbf{w}^*)} \quad (\text{I.31})$$

donde

$$\begin{aligned} P^*(c_i / \mathbf{x}_k, \mathbf{w}^*) &= \frac{p(\mathbf{x}_k / c_i, \mathbf{w}_i^*) P^*(c_i)}{\sum_{j=1}^c p(\mathbf{x}_k / c_j, \mathbf{w}_j^*) P^*(c_j)} \\ &= \frac{|C_i^*|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x}_k - \mathbf{m}_i^*)^t (C_i^*)^{-1} (\mathbf{x}_k - \mathbf{m}_i^*)\right] P^*(c_i)}{\sum_{j=1}^c |C_j^*|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x}_k - \mathbf{m}_j^*)^t (C_j^*)^{-1} (\mathbf{x}_k - \mathbf{m}_j^*)\right] P^*(c_j)} \end{aligned} \quad (\text{I.32})$$

Aunque esta notación resulta realmente compleja, su interpretación es bastante sencilla. En el caso extremo $P^*(c_i / \mathbf{x}_k, \mathbf{w}^*)$ toma el valor uno cuando \mathbf{x}_k es de la clase c_i y cero en otro caso, $P^*(c_i)$ es la fracción de muestras de c_i , \mathbf{m}_i^* es la media de esas muestras y C_i^* es la correspondiente matriz de covarianza de las muestras. De forma más general, $P^*(c_i / \mathbf{x}_k, \mathbf{w}^*)$ toma valores entre cero y uno y todas las muestras juegan algún papel en las estimas.