Clasificación entre dos poblaciones/clases

Planteamiento del Problema

Sean P_1 o C_1 y P_2 o C_2 dos poblaciones/clases donde se define una variable aleatoria vectorial X k-dimensional. Supondremos que **X** es absolutamente continua y que las funciones de densidad de ambas poblaciones, f_1 y f_2, son conocidas. Se estudia el problema de clasificar un nuevo individuo en una de estas poblaciones cuando se observa un vector $\mathbf{x_0}$. El problema puede enfocarse desde el punto de vista de la Inferencia o de la Teoría de Decisión, incluyendo además probabilidades a priori (enfoque bayesiano) o no. A continuación, se presenta la formulación del problema más general como un problema bayesiano de decisión.

Se consideran las hipótesis siguientes:
i) Las probabilidades a priori de que un individuo tomado al azar provenga de cada clase/población son conocidas: $P(C_1)$, $P(C_2)$, tales que $P(C_1)+P(C_2)=1$

ii) Las consecuencias asociadas a los errores de clasificación son $c(C_2|C_1)$ y $c(C_1|C_2)$, donde $c(C_i|C_j)$ es el coste de clasificar en C_i de un patrón que pertenece realmente a C_j. Estos costes se suponen conocidos.

iii) Las preferencias del decisor por las consecuencias de sus acciones son lineales, es decir, maximizar la función de utilidad equivale a minimizar el coste esperado o coste de oportunidad de la decisión. Por lo tanto, podemos minimizar los costes de oportunidad de la decisión mediante el criterio del valor esperado.

Las posibles decisiones en el problema son únicamente dos: asignar en C_1 ó en C_2.

Una regla de decisión equivale a hacer una partición del espacio muestral Ex (que en general será R^K) en dos regiones: A_1 y $A_2 = E_x - A_1$, tales que:

1. Si $x_0 \in A_1 \rightarrow d_1$ (asignar en C_1).
2. Si $x_0 \in A_2 \rightarrow d_2$ (asignar en C_2).

Un vez observado el valor x_0 podemos calcular la probabilidad a posteriori de que el elemento pertenezca a cada población.

Errores de clasificación

Se denomina $P(C_1|x_0)$ la probabilidad a posteriori de que un elemento que ha tomado un valor igual a x_0 pertenezca a la clase C_1. Por el teorema de Bayes esta probabilidad es:

$$P(C_1 \mid x_0) = \frac{P(x_0 \mid C_1)P(C_1)}{P(x_0 \mid C_1)P(C_1) + P(x_0 \mid C_2)P(C_2)} .$$

Las probabilidades $P(x_0|C_1)$ y $P(x_0|C_2)$ son proporcionales a las funciones de densidad de ambas poblaciones, $f_1(\mathbf{x})$ y $f_2(\mathbf{x})$ y que supondremos conocidas, por lo que la ecuación anterior puede escribirse como

$$P(C_1 \mid x_0) = \frac{f_1(x_0)P(C_1)}{f_1(x_0)P(C_1) + f_2(x_0)P(C_2)} \quad ,$$

y para la segunda clase

$$P(C_2 \mid x_0) = \frac{f_2(x_0)P(C_2)}{f_1(x_0)P(C_1) + f_2(x_0)P(C_2)} \quad , .$$

Así, si clasificamos al elemento en la clase 2 las posibles consecuencias son:

(i) Acertar con probabilidad $P(C_2|x_0)$, en cuyo caso no hay ningún coste de penalización.

(ii) Equivocarnos con probabilidad $P(C_1|x_0)$, en cuyo caso incurrimos en el coste asociado $c(C_2|C_1)$ y el coste promedio o valor esperado de la decisión "clasificar x_0 en la clase C_2" será:

$$E(d_2) = c(C_2|C_1)P(C_1|x_0)+0P(C_2|x_0) = c(C_2|C_1)P(C_1|x_0).$$

Análogamente, el coste esperado de la decisión: clasificar en la clase C_1:

$$E(d_1)=0P(C_1|x_0) + c(C_1|C_2)P(C_2|x_0) = c(C_1|C_2)P(C_2|x_0).$$

Asignaremos el elemento **x** a la clase C_2 si su coste esperado es menor, es decir, sustituyendo en las expresiones anteriores, si:

$$\frac{f_2(x_0)P(C_2)}{c(C_2 \mid C_1)} > \frac{f_1(x_0)P(C_1)}{c(C_1 \mid C_2)} \quad ,$$

que indica que, siendo iguales el resto de términos, clasificaremos en la clase C_2 si

(i) la probabilidad a priori es más alta;

(ii) la verosimilitud de que provenga de la clase C_2 es más alta;

(iii) el coste de equivocarnos al clasificarlo en la clase C_2 es más bajo.

Poblaciones Normales: Función discriminante lineal

Vamos a aplicar el análisis discriminante lineal al caso en el que las distribuciones de probabilidad asociadas a las dos clases, f_1 y f_2 sean distribuciones normales con distintos vectores de medias, pero idéntica matriz de varianzas-covarianzas. Entonces, las funciones de densidad son

$$f_i(\mathbf{x}) = (2\pi)^{-n/2} \, |\mathbf{\Sigma}|^{-1/2} \exp\left\{-1/2(\mathbf{x}-\mathbf{m}_i)^T \mathbf{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_i)\right\} \text{ para i=1,2}$$

La regla de decisión para clasificar un nuevo patrón **x** en la clase C_1 es si:

$$P(C_1)\times f_1(\mathbf{x})\times c(C_2 \mid C_1) > P(C_2)\times f_2(\mathbf{x})\times c(C_1 \mid C_2) \quad ,$$

donde $P(C_1)$ y $P(C_2)$ son las probabilidades a priori de pertenencia a cada una de las clases y $c(C_1|C_2)$ es el coste de clasificar un patrón en la clase C_1 dado que pertenece

realmente a la clase C_2. Como ambos términos son siempre positivos, tomando logaritmos y sustituyendo $f_i(x)$ por su expresión, la ecuación anterior se convierte en:

$$\ln P(C_1) - \frac{1}{2}(\mathbf{x}-\mathbf{m_1})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_1}) > \ln P(C_2) - \frac{1}{2}(\mathbf{x}-\mathbf{m_2})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_2}) - \ln\left(\frac{c(C_2/C_1)}{c(C_1/C_2)}\right)$$

es decir operando tenemos que **x** pertenece a C_1 si se verifica que

$$(\mathbf{x}-\mathbf{m_1})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_1}) < (\mathbf{x}-\mathbf{m_2})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_2}) - 2\ln\left(\frac{c(C_2/C_1)\times P(C_1)}{c(C_1/C_2)\times P(C_2)}\right) \quad (1)$$

Llamando D_i a la distancia de Mahalanobis entre el patrón observado, **x**, y la clase i:

$$D_i^2(\mathbf{x}) = (\mathbf{x}-\mathbf{m}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m}_i) \text{ para i=1,2}$$

y suponiendo iguales los costes y las probabilidades a priori, $c(C_1|C_2) = c(C_2|C_1)$; $P(C_1) = P(C_2)$, la regla resultante es:

El patrón **x** se clasifica en la clase C_1 si $D_1^2 < D_2^2$

es decir, clasificar la observación en la población de cuya media esté más próxima, usando la distancia de Mahalanobis. Observemos que si las variables **x** tuvieran matriz de covarianzas $\boldsymbol{\Sigma} = \mathbf{I}\sigma^2$, esto implica que las variables están incorreladas, la regla equivaldría a utilizar la distancia euclídea.

Interpretación de la regla de clasificación

La regla general anterior puede escribirse de una forma equivalente que permite interpretar mejor el método de clasificación utilizado.

Si simplificamos la ecuación (1)

$$(\mathbf{x}-\mathbf{m_1})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_1}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{m_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{m_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{m_1}$$

y análogamente

$$(\mathbf{x}-\mathbf{m_2})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{m_2}) = \mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - 2\mathbf{m_2}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{m_2}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_2$$

entonces, la regla divide al conjunto posible de valores de **X** en dos regiones, cuya frontera de decisión es (simplificando términos comunes en ambos miembros), la ecuación:

$$-2\mathbf{m_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{m_1}^T\boldsymbol{\Sigma}^{-1}\mathbf{m_1} = -2\mathbf{m_2}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{m_2}^T\boldsymbol{\Sigma}^{-1}\mathbf{m}_2 - 2\ln\left(\frac{c(C_2/C_1)\times P(C_1)}{c(C_1/C_2)\times P(C_2)}\right)$$

que como función de **x** equivale a:

$$(\mathbf{m_1}-\mathbf{m}_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{x} = (\mathbf{m_1}-\mathbf{m}_2)^T\boldsymbol{\Sigma}^{-1}\left(\frac{\mathbf{m_1}+\mathbf{m}_2}{2}\right) - \ln\left(\frac{c(C_2/C_1)\times P(C_1)}{c(C_1/C_2)\times P(C_2)}\right)$$

Llamando

$$\mathbf{w} = \mathbf{\Sigma}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \quad (2)$$

entonces, la frontera entre las regiones de clasificación para C_1 y C_2 puede escribirse como:

$$\mathbf{w}^T\mathbf{x} = \mathbf{w}^T(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2}) - \ln\left(\frac{c(C_2/C_1)\times P(C_1)}{c(C_1/C_2)\times P(C_2)}\right) \quad (3)$$

que es la ecuación de un hiperplano. Esta ecuación indica que el procedimiento de clasificación puede resumirse así:

(i) Calcular el vector **w** según la ecuación (2), y a continuación el segundo miembro de (3) que sólo depende de términos conocidos;

(ii) Escribir la función discriminante:

$$g(\mathbf{x})=\mathbf{w}^T\mathbf{x}=w_1x_1+w_2x_2+...+w_kx_k$$

Esta función es una combinación lineal de los valores de la variable con los pesos dados por el vector **w**.

(iii) Introducir en esta función los valores observados para el nuevo individuo a clasificar, $\mathbf{x}_0=(x_{10},...,x_{k0})$. Según la ecuación (1) clasificaremos en la clase 1 cuando el primer miembro sea mayor que el segundo. En el caso particular de que $c(C_1|C_2)\times P(C_2) = c(C_2|C_1)\times P(C_1)$ la regla de decisión se reduce entonces a clasificar en C_1 si:

$$\mathbf{w}^T\mathbf{x} < \mathbf{w}^T(\frac{\mathbf{m}_1 + \mathbf{m}_2}{2})$$

Se puede comprobar que esta regla equivale a proyectar el punto **x** que queremos clasificar y las medias de ambas poblaciones sobre una recta, y después asignar el punto a aquella población de cuya media se encuentre más próxima en la proyección. En resumen, el problema de clasificación cuando los costes y las probabilidades a priori se suponen idénticos y las variables normales, se reduce a definir una variable escalar, $z = \mathbf{w}^T\mathbf{x}$, trasladar las medias y el punto observado a dicha escala, y asignarlo a la media más próxima. La distancia entre las medias proyectadas es igual a su distancia de Mahalanobis en el espacio. La varianza de la nueva variable escalar, z, es igual a la distancia de Mahalanobis entre las medias

$$Var(z) = Var(\mathbf{w}^T\mathbf{x}) = \mathbf{w}^T Var(\mathbf{x})\mathbf{w} = \mathbf{w}^T\mathbf{\Sigma}\mathbf{w} = (\mathbf{m}_1 - \mathbf{m}_2)^T\mathbf{\Sigma}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) = D^2$$

ya que $\mathbf{w} = \mathbf{\Sigma}^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$