



**FACULTAD de
CIENCIAS ECONÓMICAS**

EL CLASIFICADOR BAYESIANO ÓPTIMO

ALAN REYES-FIGUEROA

ELEMENTS OF MACHINE LEARNING

(AULA 20) 26.ABRIL.2024

Función de costo

Dado un clasificador $\hat{y} = g : S \rightarrow \Omega$, definimos una **función de pérdida** o **costo** $L(y, \hat{y}(\mathbf{x}))$.

Ejemplos:

- $L(y, \hat{y}(\mathbf{x})) = \begin{cases} 1, & \text{si } y \neq \hat{y}(\mathbf{x}); \\ 0, & \text{si } y = \hat{y}(\mathbf{x}). \end{cases}$
- $L(y, \hat{y}(\mathbf{x})) = (y - \hat{y}(\mathbf{x}))^2.$
- $L(y, \hat{y}(\mathbf{x})) = \begin{cases} c_1, & \text{si } y = 1, \hat{y}(\mathbf{x}) = 0; \\ c_2, & \text{si } y = 0, \hat{y}(\mathbf{x}) = 1; \\ c_3, & \text{si } y = \hat{y}(\mathbf{x}). \end{cases}$
- $L(y, \hat{y}(\mathbf{x})) = |y - \hat{y}(\mathbf{x})|.$
- $L(y, \hat{y}(\mathbf{x})) = \log \cosh (y - \hat{y}(\mathbf{x})).$

Definimos el **error de clasificación** como $\mathbb{E}(L(y, \hat{y}(\mathbf{x})))$. El **error empírico** se define como

$$\frac{1}{n} \sum_i L(y_i, \hat{y}(\mathbf{x}_i)).$$

Clasificador bayesiano óptimo

Típicamente, la función de costo satisface $L(y, \hat{y}(\mathbf{x})) \geq 0$.

Dado un conjunto de datos $(X, Y) \sim \mathbb{P}$, y una función de costo $L \geq 0$, queremos encontrar un clasificador $\hat{y}(\mathbf{x})$ tal que

$$\mathbb{E}(L(Y, \hat{Y}(X))) = \mathbb{E}_{X,Y}(L(Y, \hat{Y}(X))) \text{ sea mínima.} \quad (1)$$

En otros casos, nos puede interesar minimizar la probabilidad $\mathbb{P}(\sum_i L(y_i, \hat{y}(\mathbf{x}_i)) > \text{threshold})$.

De (1)

$$\begin{aligned} \mathbb{E}_{X,Y}(L(Y, \hat{Y}(X))) &= \mathbb{E}_X \mathbb{E}_{Y|X=\mathbf{x}}(L(Y, \hat{Y}(X))) \\ &= \int_{\mathbb{R}^d} \mathbb{E}_{Y|X=\mathbf{x}} L(Y, \hat{Y}(\mathbf{x})) f_X(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (2)$$

Clasificador bayesiano óptimo

La ecuación en (2) es importante porque de alguna manera indica que el problema de minimización es desacoplado: se puede minimizar de forma separada en X y se puede también minimizar en Y .

Si minimizamos lo anterior sobre $\hat{Y}$, es suficiente para cada $\mathbf{x}$ minimizar la siguiente función

$$\operatorname{argmin}_{\hat{Y}(\mathbf{x})} \mathbb{E}_{Y|X=\mathbf{x}} L(Y, \hat{Y}(\mathbf{x})), \quad \forall \mathbf{x}.$$

Definición

La solución a la ecuación anterior se llama el **clasificador bayesiano óptimo**.

Obs! Este es un clasificador teórico. En la práctica no es posible calcularlo, a menos que dispongamos de la información de la distribución conjunta $\mathbb{P}(X, Y)$ teórica.

Se utiliza para dar estimados teóricos y cotas del error de clasificación.

Clasificador bayesiano óptimo

Observaciones:

- Se llama *óptimo* porque es lo mejor que podemos hacer en el caso que tenemos la información completa $\mathbb{P}_{X,Y}$.
- En el caso finito, la integral $\mathbb{E}_{X,Y}(L(Y, \hat{Y}(X)))$ se reduce a una suma.
- Se puede mostrar que en el caso finito, el clasificador bayesiano óptimo es la asignación $\hat{y}$ que minimiza la “probabilidad de cometer un error”

$$n \mathbb{E}_{X,Y}(L(Y, \hat{Y}(X))) = \sum_y \sum_{\mathbf{x}: y(\mathbf{x})=y} L(y, \hat{y}(\mathbf{x})) \mathbb{P}(y \neq \hat{y}(\mathbf{x})).$$

Clasificador bayesiano óptimo

Ejemplo: $Y \sim \text{Ber}(p)$

Si Y toma solamente dos valores, 0 y 1, entonces podemos escribir

$$\mathbb{E}_{Y|X=\mathbf{x}} L(Y, \hat{y}(\mathbf{x})) = L(0, \hat{y}(\mathbf{x})) \mathbb{P}(Y = 0 | X = \mathbf{x}) + L(1, \hat{y}(\mathbf{x})) \mathbb{P}(Y = 1 | X = \mathbf{x}).$$

Denotemos por $\lambda_{ij} = L(i, j) = L(y = i, \hat{y} = j)$, para $i, j \in \{0, 1\}$.

Entonces si tomamos el caso binario y el costo de un falso positivo igual a un falso negativo (costo simétrico):

$$\lambda_{00} = L(0, 0) = 0, \quad \lambda_{11} = L(1, 1) = 0, \quad \lambda_{01} = L(0, 1) = 1 = \lambda_{10} = L(1, 0),$$

entonces tenemos los costos

$$\text{si } \hat{y}(\mathbf{x}) = 0 : \quad L(0, \hat{y}(\mathbf{x})) = \lambda_{00} = 0, \quad L(1, \hat{y}(\mathbf{x})) = \lambda_{10} = 1,$$

$$\text{si } \hat{y}(\mathbf{x}) = 1 : \quad L(0, \hat{y}(\mathbf{x})) = \lambda_{01} = 1, \quad L(1, \hat{y}(\mathbf{x})) = \lambda_{11} = 0.$$

Clasificador bayesiano óptimo

y el error sería

$$\begin{aligned} \text{si } \hat{y}(\mathbf{x}) = 0 : & \quad \text{el error es } \mathbb{P}(Y = 1 \mid X = \mathbf{x}), \\ \text{si } \hat{y}(\mathbf{x}) = 1 : & \quad \text{el error es } \mathbb{P}(Y = 0 \mid X = \mathbf{x}). \end{aligned}$$

Así, el clasificador bayesiano óptimo es

$$\hat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } \mathbb{P}(Y = 0 \mid X = \mathbf{x}) > \mathbb{P}(Y = 1 \mid X = \mathbf{x}) \\ 1, & \text{si } \mathbb{P}(Y = 1 \mid X = \mathbf{x}) > \mathbb{P}(Y = 0 \mid X = \mathbf{x}) \end{cases} \quad (3)$$

Obs! En este caso, $\hat{y}$ asigna $\mathbf{x}$ a la categoría más probable según $\mathbb{P}(Y \mid X = \mathbf{x})$.

Clasificador bayesiano óptimo

Podemos aún simplificar esto usando la regla de Bayes. Escribimos

$$\begin{aligned}(3) \quad &\Longleftrightarrow \mathbb{P}(Y = 0 \mid X = \mathbf{x}) > \mathbb{P}(Y = 1 \mid X = \mathbf{x}) \\&\Longleftrightarrow \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 0) \mathbb{P}(Y = 0)}{\mathbb{P}(X = \mathbf{x})} > \frac{\mathbb{P}(X = \mathbf{x} \mid Y = 1) \mathbb{P}(Y = 1)}{\mathbb{P}(X = \mathbf{x})} \\&\Longleftrightarrow \mathbb{P}(X = \mathbf{x} \mid Y = 0) \mathbb{P}(Y = 0) > \mathbb{P}(X = \mathbf{x} \mid Y = 1) \mathbb{P}(Y = 1) \\&\Longleftrightarrow f_0(\mathbf{x}) \mathbb{P}(Y = 0) > f_1(\mathbf{x}) \mathbb{P}(Y = 1);\end{aligned}$$

donde la $f_i(\mathbf{x})$ representa la función de densidad o masa de probabilidad condicional

$$f_i(\mathbf{x}) = \mathbb{P}(X = \mathbf{x} \mid Y = i).$$

Clasificador bayesiano óptimo

En la ecuación anterior

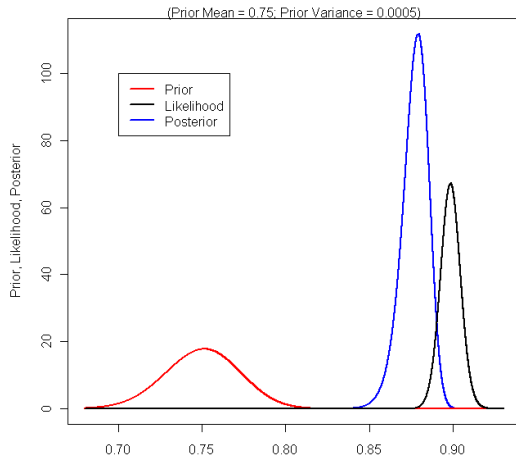
- $f_i(\mathbf{x}) = \mathbb{P}(X = \mathbf{x} \mid Y = i)$ representa la verosimilitud,
- $\mathbb{P}(Y = i)$ representa la distribución o información *a priori* de la categoría $Y = i$,
- mientras que el cociente en la regla de Bayes

$$\frac{f_i(\mathbf{x}) \mathbb{P}(Y = i)}{\mathbb{P}(X = \mathbf{x})}$$

representa la probabilidad *a posteriori*.

Entonces, el clasificador Bayesiano óptimo $\hat{y}$ asigna $\mathbf{x}$ a la categoría más probable según la probabilidad a posterior.

Clasificador bayesiano óptimo



La distribución posterior es una mezcla entre la previa y la verosimilitud.

Clasificador bayesiano óptimo

Ejemplo: (Caso general con costo 0 al clasificar correcto)

El error es

$$\begin{aligned}\text{si } \hat{y}(\mathbf{x}) = 0 : & \quad \text{el error es } \lambda_{10}\mathbb{P}(Y = 1 \mid X = \mathbf{x}), \\ \text{si } \hat{y}(\mathbf{x}) = 1 : & \quad \text{el error es } \lambda_{01}\mathbb{P}(Y = 0 \mid X = \mathbf{x}).\end{aligned}$$

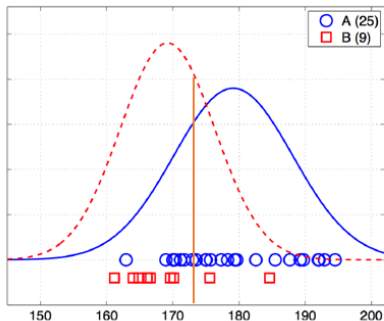
Así, el clasificador bayesiano óptimo es

$$\hat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } \lambda_{01}\mathbb{P}(Y = 0 \mid X = \mathbf{x}) > \lambda_{10}\mathbb{P}(Y = 1 \mid X = \mathbf{x}) \\ 1, & \text{si } \lambda_{10}\mathbb{P}(Y = 1 \mid X = \mathbf{x}) > \lambda_{01}\mathbb{P}(Y = 0 \mid X = \mathbf{x}) \end{cases} \quad (4)$$

Usando la regla de Bayes, obtenemos

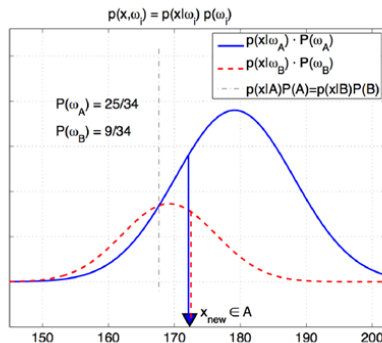
$$\hat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } \lambda_{01}f_0(\mathbf{x})\mathbb{P}(Y = 0) > \lambda_{10}f_1(\mathbf{x})\mathbb{P}(Y = 1) \\ 1, & \text{si } \lambda_{10}f_1(\mathbf{x})\mathbb{P}(Y = 1) > \lambda_{01}f_0(\mathbf{x})\mathbb{P}(Y = 0) \end{cases} \quad (5)$$

Clasificador bayesiano óptimo



Maximum Likelihood classifier

Predicted class: **Female**



Bayes Classifier

Predicted class: **Male**

Clasificador bayesiano óptimo

Ejemplo: (El caso general)

El error es

si $\hat{y}(\mathbf{x}) = 0$: el error es $\lambda_{00}\mathbb{P}(Y = 0 | X = \mathbf{x}) + \lambda_{10}\mathbb{P}(Y = 1 | X = \mathbf{x})$,

si $\hat{y}(\mathbf{x}) = 1$: el error es $\lambda_{01}\mathbb{P}(Y = 0 | X = \mathbf{x}) + \lambda_{11}\mathbb{P}(Y = 1 | X = \mathbf{x})$.

Así, el clasificador bayesiano óptimo es

$$\hat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } (\lambda_{00} - \lambda_{01})\mathbb{P}(Y = 0 | X = \mathbf{x}) < (\lambda_{11} - \lambda_{10})\mathbb{P}(Y = 1 | X = \mathbf{x}); \\ 1, & \text{si } (\lambda_{11} - \lambda_{10})\mathbb{P}(Y = 1 | X = \mathbf{x}) < (\lambda_{00} - \lambda_{01})\mathbb{P}(Y = 0 | X = \mathbf{x}) \end{cases}$$

Usando la regla de Bayes, obtenemos

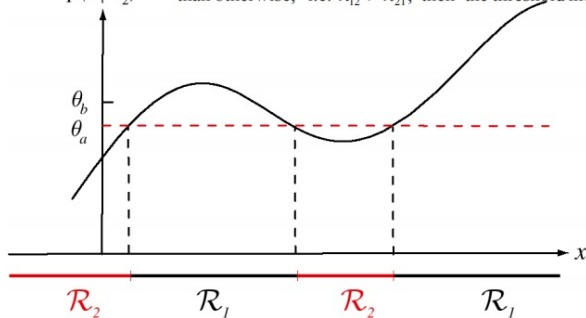
$$\hat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } (\lambda_{00} - \lambda_{01})f_0(\mathbf{x})\mathbb{P}(Y = 0) < (\lambda_{11} - \lambda_{10})f_1(\mathbf{x})\mathbb{P}(Y = 1); \\ 1, & \text{si } (\lambda_{11} - \lambda_{10})f_1(\mathbf{x})\mathbb{P}(Y = 1) < (\lambda_{00} - \lambda_{01})f_0(\mathbf{x})\mathbb{P}(Y = 0) \end{cases} \quad (6)$$

Clasificador bayesiano óptimo

$$\frac{P(\mathbf{x} | \omega_1)}{P(\mathbf{x} | \omega_2)} > \frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda$$

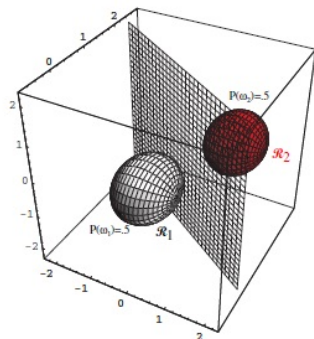
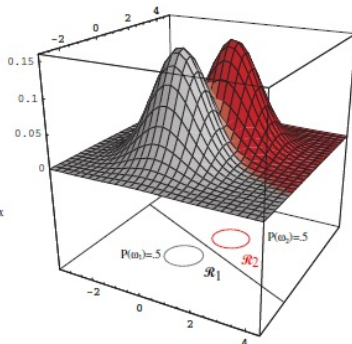
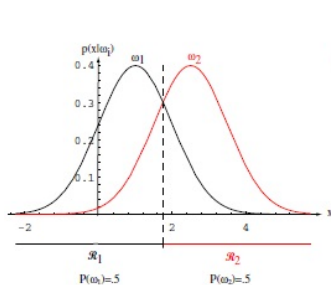
$$\frac{p(x|\omega_1)}{p(x|\omega_2)}$$

If misclassifying ω_2 as ω_1 becomes more expensive than otherwise, i.e. $\lambda_{12} > \lambda_{21}$, then the threshold increases



Regla de decisión para el clasificador Bayesiano óptimo.

Ejemplos



Fronteras de decisión del clasificador bayesiano óptimo, para el caso de dos normales $f_i(\mathbf{x})$.

Ejemplos

