El Clasificador Bayesiano Óptimo

Alan Reyes-Figueroa

Introducción a la Ciencia de Datos (Aula 29) 09.mayo.2022

Función de costo

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

Ejemplos:

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

Definimos el **error de clasificación** como $\mathbb{E}\big(L(y, \widehat{y}(\mathbf{x}))\big)$. El **error empírico** se define como $\frac{1}{n}\sum_i L(y_i, \widehat{y}(\mathbf{x}_i))$.

Clasificador bayesiano óptimo

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

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

$$\mathbb{E}\big(L(Y, \widehat{Y}(X))\big) = \mathbb{E}_{X,Y}\big(L(Y, \widehat{Y}(X))\big) \;\; \text{sea mínima}. \tag{1}$$

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

De (1)

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

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 $\widehat{Y}$, es suficiente para cada $\mathbf{x}$ minimizar la siguiente función

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

Definición

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

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 $\mathbb{E}_{X,Y}\big(L(Y,\widehat{Y}(X))\big)$ se reduce a una suma.

- Se puede mostrar que en el caso finito, el clasificador bayesiano óptimo es la asignación $\widehat{y}$ que minimiza la "probabilidad de cometer un error"

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

UNIVERSIDAD DEL VALLE DE GUATEMALA

Clasificador bayesiano óptimo

Ejemplo: $Y \sim Ber(p)$

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

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

Denotemos por $\lambda_{ij} = L(i, j) = L(y = i, \widehat{y} = j)$, 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

$$\begin{aligned} \text{si } \widehat{y}(\mathbf{x}) = 0: &\quad L(0, \widehat{y}(\mathbf{x})) = \lambda_{00} = 0,\ L(1, \widehat{y}(\mathbf{x})) = \lambda_{10} = 1, \\ \text{si } \widehat{y}(\mathbf{x}) = 1: &\quad L(0, \widehat{y}(\mathbf{x})) = \lambda_{01} = 1,\ L(1, \widehat{y}(\mathbf{x})) = \lambda_{11} = 0. \end{aligned}$$

Clasificador bayesiano óptimo

y el error sería

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

Así, el clasificador bayesiano óptimo es

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

Obs! En este caso, $\widehat{y}$ 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

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

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

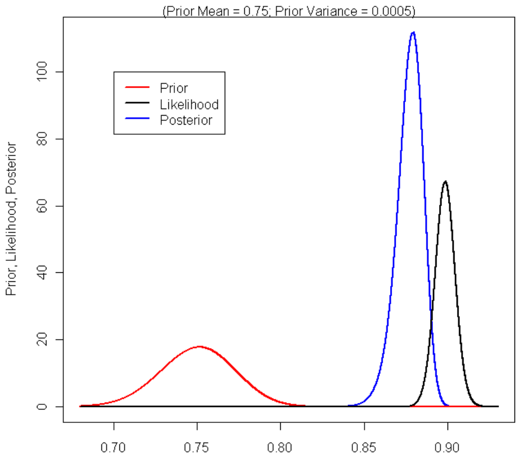
- $f_i(\mathbf{x}) = \mathbb{P}(X = \mathbf{x} \mid Y = i)$ representa la verosimilitud,
- $\mathbb{P}(Y = i)$ representa la distribución o información *a priori* de la categoría $Y = i$,
- mientras que el cociente en la regla de Bayes

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

  representa la probabilidad *a posteriori*.

Entonces, el clasificador Bayesiano óptimo $\widehat{y}$ 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

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

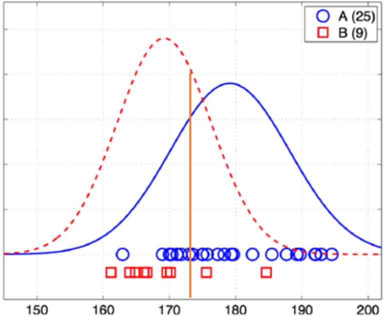
Así, el clasificador bayesiano óptimo es

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

Usando la regla de Bayes, obtenemos

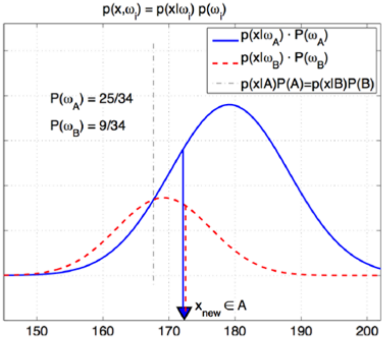
$$\widehat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } \lambda_{01}f_0(\mathbf{x})\,\mathbb{P}(Y = 0) > \lambda_{10}f_1(\mathbf{x})\,\mathbb{P}(Y = 1) \\ 1, & \text{si } \lambda_{10}f_1(\mathbf{x})\,\mathbb{P}(Y = 1) > \lambda_{01}f_0(\mathbf{x})\,\mathbb{P}(Y = 0) \end{cases} \tag{5}$$

Clasificador bayesiano óptimo



Maximum Likelihood classifier

Predicted class: **Female**



Bayes Classifier

Predicted class: **Male**

UNIVERSIDAD DEL VALLE DE GUATEMALA

Clasificador bayesiano óptimo

Ejemplo: (El caso general)

El error es

$$\begin{aligned} &\text{si } \widehat{y}(\mathbf{x}) = 0: &&\text{el error es } \lambda_{00}\mathbb{P}(Y=0 \mid X=\mathbf{x}) + \lambda_{10}\mathbb{P}(Y=1 \mid X=\mathbf{x}),\\ &\text{si } \widehat{y}(\mathbf{x}) = 1: &&\text{el error es } \lambda_{01}\mathbb{P}(Y=0 \mid X=\mathbf{x}) + \lambda_{11}\mathbb{P}(Y=1 \mid X=\mathbf{x}). \end{aligned}$$

Así, el clasificador bayesiano óptimo es

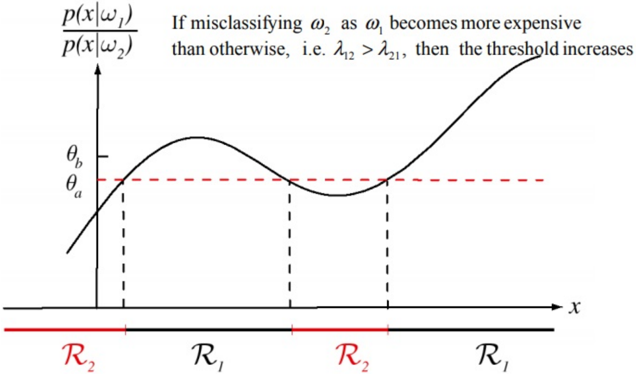
$$\widehat{y}(\mathbf{x}) = \begin{cases} 0, & \text{si } (\lambda_{00} - \lambda_{01})\mathbb{P}(Y=0 \mid X=\mathbf{x}) < (\lambda_{11} - \lambda_{10})\mathbb{P}(Y=1 \mid X=\mathbf{x});\\ 1, & \text{si } (\lambda_{11} - \lambda_{10})\mathbb{P}(Y=1 \mid X=\mathbf{x}) < (\lambda_{00} - \lambda_{01})\mathbb{P}(Y=0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

Usando la regla de Bayes, obtenemos

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

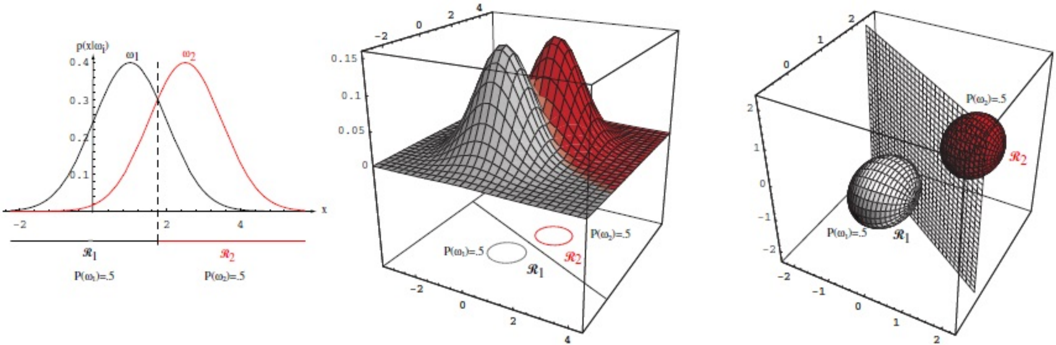
Clasificador bayesiano óptimo

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



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