

Bonus 2

Luis Suárez Lloréns

Regresión logística

Cuestión. 4. Consideremos el caso de la verificación de la huella digital (ver transparencias de clase). Tras aprender con un modelo de regresión logística a partir de datos obtenemos una función una hipótesis final

$$g(x) = \mathbb{P}[y = +1|\mathbf{x}]$$

que representa la estimación de la probabilidad de que $y = +1$. Suponga que la matriz de coste está dada por

		Verdadera Clasificación	
		+1 (persona correcta)	-1 (intruso)
decisión	+1	0	c_a
decisión	-1	c_r	0

Para una nueva persona con huella digital \mathbf{x} , calculamos $g(\mathbf{x})$ y tenemos que decidir si aceptar o rechazar a la persona (i.e. tenemos que usar una decisión 1/0). Por tanto aceptaremos si $g(\mathbf{x}) \geq \kappa$, donde κ es un umbral.

1. Definir la función de costo(aceptar) como el costo esperado si se acepta la persona. Definir de forma similar el costo(rechazo). Mostrar que

$$\begin{aligned}\text{costo(aceptar)} &= (1 - g(\mathbf{x}))c_a \\ \text{costo(rechazar)} &= g(\mathbf{x})c_r\end{aligned}$$

2. Usar el apartado anterior para derivar una condición sobre $g(x)$ para aceptar la persona y mostrar que

$$\kappa = \frac{c_a}{c_a + c_r}$$

3. Usar las matrices de costo para la aplicación del supermercado y la CIA (transparencias de clase) para calcular el umbral κ para cada una de las dos clases. Dar alguna interpretación del umbral obtenido.

Respuesta.

1. Vamos a ver las funciones costo suponiendo que hemos aceptado o que hemos rechazado a la persona. Si aceptamos correctamente, la función costo vale 0, y la probabilidad de que esto pase es $g(x)$. Si aceptamos incorrectamente, la función de coste vale c_a y la probabilidad es la contraria a aceptar correctamente, $1 - g(x)$. Por tanto, el costo esperado de aceptar es:

$$\text{costo(aceptar)} = 0g(x) + c_a(1 - g(x)) = (1 - g(x))c_a$$

De manera similar, hacemos costo(rechazar) :

$$\text{costo(rechazar)} = c_r g(x) + 0(1 - g(x)) = g(x)c_r$$

2. Para seleccionar κ , vemos cuando se igualan los costes de aceptar y rechazar. Substituyendo en las ecuaciones del costo anteriores $g(x)$ por κ se obtiene:

$$(1 - \kappa)c_a = \kappa c_r \rightarrow c_a = \kappa(c_r + c_a) \rightarrow \kappa = \frac{c_a}{c_a + c_r}$$

3. En el caso del supermercado, tenemos que $c_a = 1$ y $c_r = 10$. Entonces $\kappa = \frac{1}{11} \simeq 0.09$. Por otro lado, en el caso de la CIA, tenemos que $c_a = 1000$ y $c_r = 1$ y $\kappa = \frac{1000}{1001} \simeq 0.999$.

Esto nos indica que en el caso de la CIA aceptamos a un usuario si la probabilidad de que sea la persona correcta es mayor del 99.9%. Esto es normal, pues la seguridad de la CIA es muy alta y sólo permite pasar si se está muy seguro de que es la persona correcta.

En el caso de un supermercado, se puede ser muy permisivo y queremos que el sistema de detección no supongan un impedimento para los usuarios. Por eso, se acepta cuando hay un 10% o más de probabilidad de que sea la persona correcta. Es decir, sólo se impide el acceso cuando estamos bastante seguros de que no es la persona correcta.