Clasificación

Verónica E. Arriola-Rios

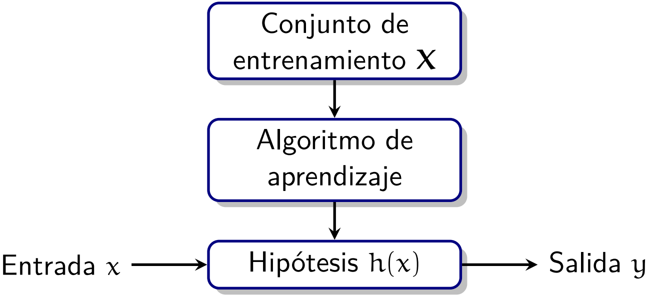
Aprendizaje de máquina

17 de agosto de 2021

Problema

- En un problema de clasificación se pretende, dadas las características de un ejemplar, identificar la clase a la cual pertenece.

Clasificación



$$h(x) : \mathbb{R}^n \to \{e | e \text{ es una etiqueta de clase}\} \tag{1}$$

Temas

Espacio de hipótesis

- Convierte las rectas en fronteras entre regiones correspondientes a clases.

$$h_\Theta(X) = \frac{1}{1 + e^{-\Theta^T X}} \tag{2}$$

Regresión logística

$$h_\Theta(X) = g(\Theta^T X) \tag{3}$$

$$g(Z) = \frac{1}{1 + e^{-Z}} \tag{4}$$

donde $g(Z)$ es la función *logística* o *sigmoide*. Obsérvese que $Z = \Theta^T X$.

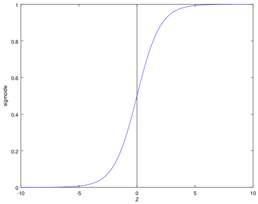


Figura: Función logística.

Regresión logística multivariada (Ejemplo)

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 \tag{5}$$

$$h_\Theta(x) = g(z) \tag{6}$$

- si $z > 0$ el ejemplar es positivo.
- si $z < 0$ el ejemplar es negativo.

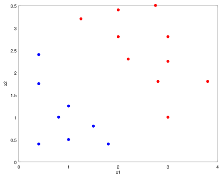


Figura: Problema de clasificación, ilustrado en el plano.

Regresión logística multivariada (Ejemplo)

$$z = \theta_0 + \theta_1 x_1^2 + \theta_2 x_2^2 \tag{7}$$

$$h_\Theta(x) = g(z) \tag{8}$$

- si $z > 0$ el ejemplar es positivo.
- si $z < 0$ el ejemplar es negativo.

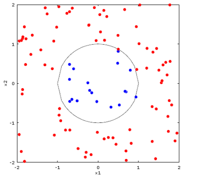


Figura: Problema de clasificación, ilustrado en el plano.

Temas

Función de costo

- La función de error cuadrático tiene numerosos máximos y mínimos locales, por lo que es difícil de optimizar.
- Se utiliza en su lugar la entropía cruzada, que para este caso es un poco más regular.

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))] \tag{9}$$

$$h_\theta(X) = \frac{1}{1+e^{-\theta^T X}} \tag{10}$$

donde m es el número de ejemplares de entrenamiento.

- Se optimiza con algoritmos numéricos, como descenso por el gradiente.

La derivada de la hipótesis

$$\frac{\partial h}{\partial \theta_i} = -\frac{1}{(1+e^{-\Theta^T X})^2} e^{-\Theta^T X}(-x_i) \tag{11}$$

$$= \frac{1}{1+e^{-\Theta^T X}} \frac{e^{-\Theta^T X}}{1+e^{-\Theta^T X}}(x_i) \tag{12}$$

$$= h_\Theta(X) \frac{e^{-\Theta^T X}}{1+e^{-\Theta^T X}}(x_i) \tag{13}$$

$$= h_\Theta(X) \frac{-1+1+e^{-\Theta^T X}}{1+e^{-\Theta^T X}}(x_i) \tag{14}$$

$$= h_\Theta(X) \left(\frac{1+e^{-\Theta^T X}}{1+e^{-\Theta^T X}} - \frac{1}{1+e^{-\Theta^T X}} \right)(x_i) \tag{15}$$

$$\frac{\partial h}{\partial \theta_i} = h_\Theta(X)(1-h_\Theta(X))x_i \tag{16}$$

Función de costo (entropía cruzada)

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Costo}(h_\theta(x^{(i)}), y^{(i)}) \tag{17}$$

$$\mathrm{Costo}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{si } y = 1 \\ -\log(1 - h_\theta(x)) & \text{si } y = 0 \end{cases} \tag{18}$$

$$= -y\log(h_\theta(x)) - (1-y)\log(1 - h_\theta(x)) \tag{19}$$

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m} y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))\right] \tag{20}$$

†$J(\theta)$ se elige utilizando máxima verosimilitud.

Función de costo y su derivada

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))\right] \tag{21}$$

$$h_\Theta(X) = \frac{1}{1+e^{-\Theta^T X}} \tag{22}$$

$$\frac{\partial J}{\partial \theta_i} = \frac{1}{m}\sum_{i=1}^{m}\left[h_\Theta(X) - y\right] x_i \tag{23}$$

Entonces, el algoritmo de descenso por el gradiente luce igual que para regresión lineal, excepto por la forma de $h_\Theta(X)$.

Derivada de la función de costo

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}[y\log(h_\theta(x)) + (1-y)\log(1-h_\theta(x))] \tag{24}$$

$$\frac{\partial J}{\partial \theta_i} = -\frac{1}{m}\sum_{i=1}^{m}\left[y\frac{\partial}{\partial \theta_i}\log(h_\theta(x)) + (1-y)\frac{\partial}{\partial \theta_i}\log(1-h_\theta(x))\right] \tag{25}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y\frac{1}{h_\theta(x)}\frac{\partial h}{\partial \theta_i} - (1-y)\frac{1}{1-h_\theta(x)}\frac{\partial h}{\partial \theta_i}\right] \tag{26}$$

$$\tag{27}$$

$$= -\frac{1}{m}\sum_{i=1}^{m}\left[y\frac{h_\Theta(X)(1-h_\Theta(X))}{h_\theta(x)} - (1-y)\frac{h_\Theta(X)(1-h_\Theta(X))}{1-h_\theta(x)}\right]x_i \tag{28}$$

$$\frac{\partial J}{\partial \theta_i} = -\frac{1}{m}\sum_{i=1}^{m}\left[y(1-h_\Theta(X)) - (1-y)h_\Theta(X)\right]x_i \tag{29}$$

$$\frac{\partial J}{\partial \theta_i} = -\frac{1}{m}\sum_{i=1}^{m}\left[y - \cancelto{0}{yh_\Theta(X)} - h_\Theta(X) + \cancelto{0}{yh_\Theta(X)}\right]x_i \tag{30}$$

$$\frac{\partial J}{\partial \theta_i} = \frac{1}{m}\sum_{i=1}^{m}\left[h_\Theta(X) - y\right]x_i \tag{31}$$

En forma vectorizada:

$$\nabla J = \frac{1}{m}X^T(H_\Theta - Y) \tag{32}$$

Temas

Matriz de confusión binaria

- Cuando se trata de evaluar a un clasificador, la función de error se usa para entrenar, pero no necesariamente sirve para indicar si el clasificador funciona correctamente o no.
- La *matriz de confusión* contabiliza los aciertos y errores efectivos que ha cometido el clasificador sobre un conjunto de ejemplares.

Tabla: Matriz de confusión

	Predicción negativa	Predicción positiva
Casos negativos	VN	FP
Casos positivos	FN	VP

Precisión y recuperación

- En inglés se les conoce como *precision and recall*.

$$P = \frac{VP}{VP + FP} \tag{33}$$

$$R = \frac{VP}{VP + FN} \tag{34}$$

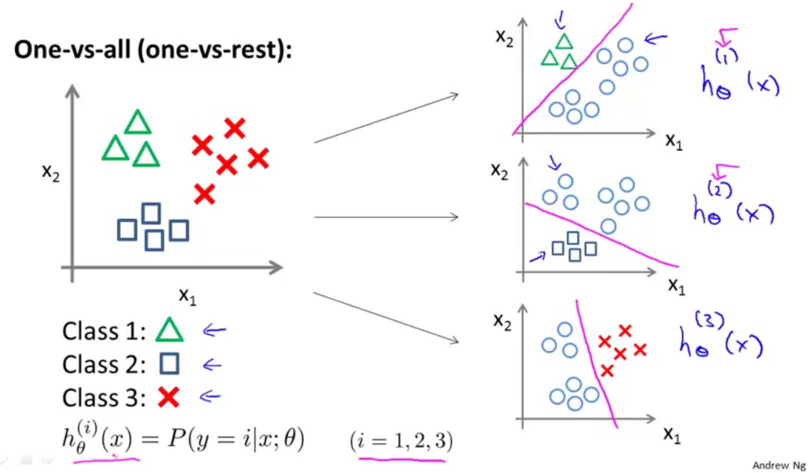
- Como resumen se utiliza la exactitud (*accurracy*)

$$A = \frac{VP + VN}{VP + VN + FP + FN} = \frac{VP + VN}{TOTAL} \tag{35}$$

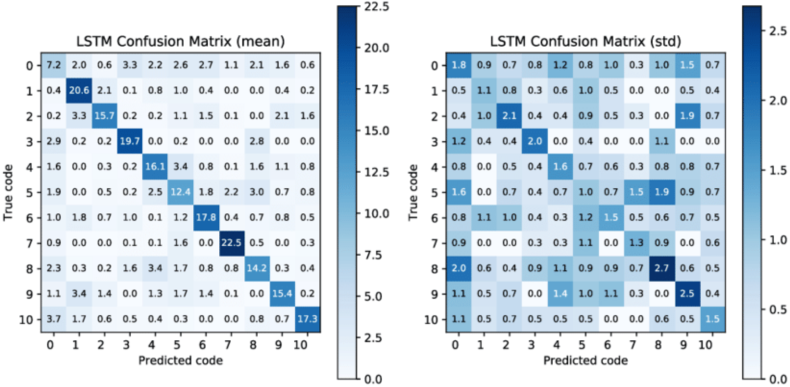
- También se utiliza la media armónica de la precisión y la recuperación, conocida como F

$$F = 2 \cdot \frac{\text{precisión} \times \text{recuperación}}{\text{precisión} + \text{recuperación}} \tag{36}$$

Uno vs. todos


One-vs-all (one-vs-rest):
x_2
x_1
Class 1: △
Class 2: □
Class 3: ✕
$h_\theta^{(i)}(x) = P(y = i|x;\theta)$
$(i = 1, 2, 3)$
$h_\theta^{(1)}(x)$
$h_\theta^{(2)}(x)$
$h_\theta^{(3)}(x)$
Andrew Ng

Matriz de confusión multiclase



Referencias I