Regresión Lineal, Logística

Inteligencia Artificial

UNIVERSIDAD EAFIT®

Notación matemática

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

Vector de características

$$\mathbf{x}^{(i)} = \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix} \in \Re^{d\times 1}$$

$$y^{(i)} \in \Re$$

$$y^{(i)} \in [0, 1, \cdots, C]$$

Matriz de características

$$\mathbf{X} = \begin{bmatrix} \cdots & \mathbf{x}^{(1)\top} & \cdots \\ \cdots & \mathbf{x}^{(2)\top} & \cdots \\ & \vdots & \\ \cdots & \mathbf{x}^{(n)\top} & \cdots \end{bmatrix} \in \Re^{n\times d} \quad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{bmatrix}$$

Métodos de aprendizaje de máquina

Formas de aproximar la relación entre $\mathbf{x}^{(i)}$ y $\mathrm{y}^{(i)}$

$$f(\mathbf{x}^{(i)}) = y^{(i)}$$

Función de costo

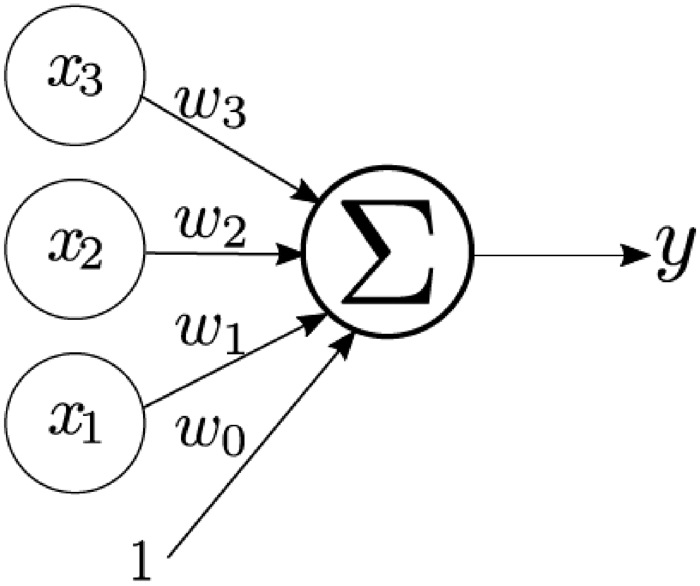
$$\mathcal{L}\left(f(\mathbf{x}^{(i)}), y^{(i)}\right)$$

Regresión Lineal

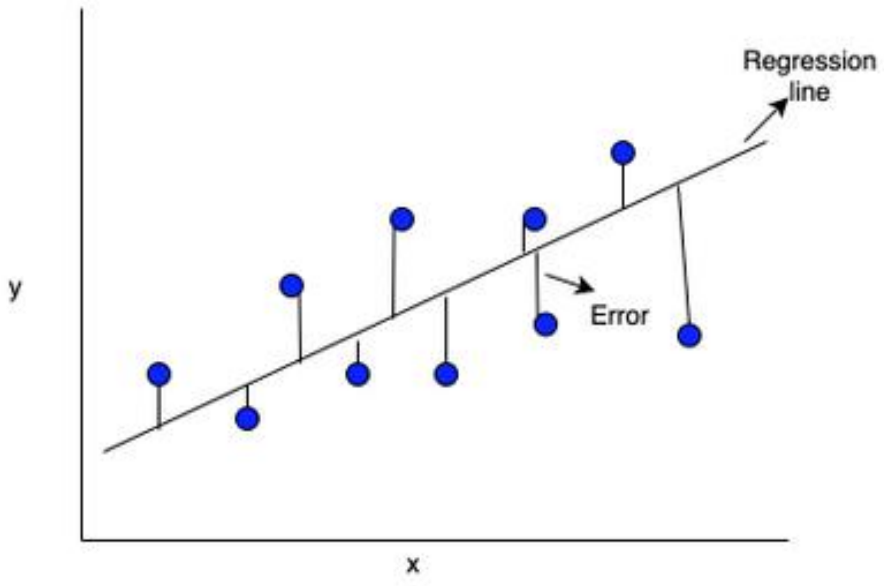
Regresión Lineal: Modelo

$$\hat{y} = \hat{f}(\mathbf{x}) = w_0 + \sum_{j=1}^{d} w_j \times x_j$$

$$\hat{y} = \mathbf{w}^T \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$$



Regresión Lineal: Error en Función Objetivo



Regresión Lineal: Función Objetivo

$$Cost_{\mathbf{w}} = \frac{1}{m} \sum_{i=1}^{n} \left(y_i - \mathbf{w}^T \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} \right)^2$$

Regresión Lineal: Entrenamiento

Descenso por el gradiente:

$$\mathbf{w} := \mathbf{w} - \alpha \times \nabla Cost_{\mathbf{w}}$$

$$\nabla Cost_{\mathbf{w}} = \frac{1}{m} X^T (\hat{y} - y)$$

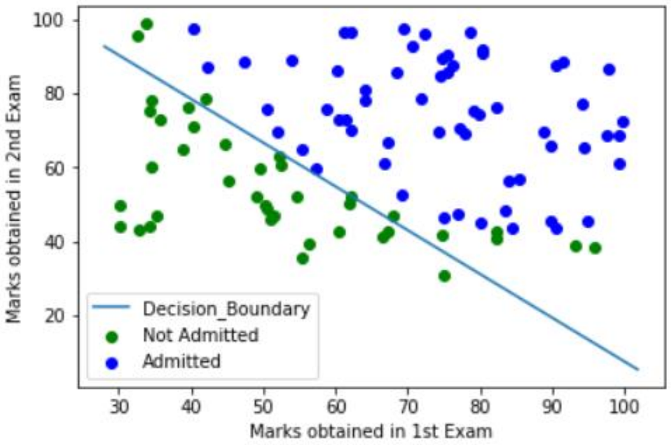
Ecuación normal:

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T Y)$$

Regresión Logística

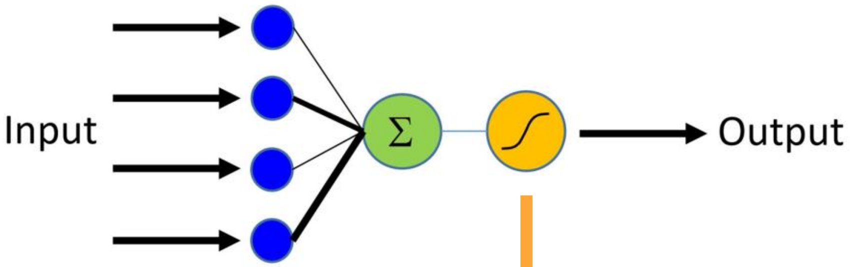
Regresión Logística: Nombre

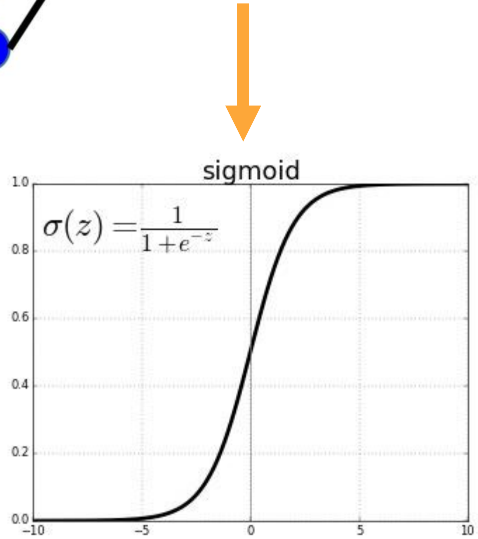
La regresión logística es un tipo de modelo lineal de **clasificación** a pesar de llevar históricamente la palabra regresión en su nombre.



Regresión Logística: Modelo

$$\hat{y} = h(\mathbf{x}) = \frac{1}{1 + \exp\left(-\mathbf{w}^T \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}\right)}$$



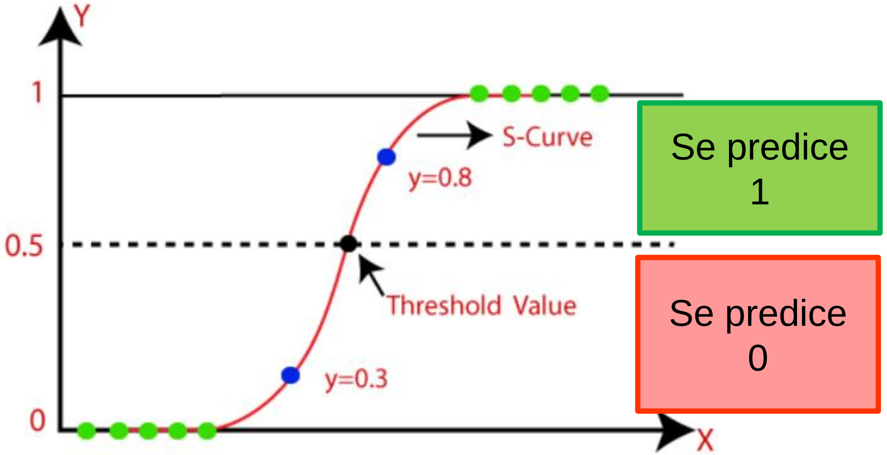


Regresión Logística: Etiqueta

La **etiqueta** con la que se entrena una regresión logística debe ser **binaria**, tomando únicamente los valores 0 o 1. La clase 0 se suele denominar **clase negativa**, mientras que la clase 1 se suele llamar **clase positiva**.

Regresión Logística: Predicción

La salida que retorna una regresión logística es un número entre 0 y 1 que se interpreta como la probabilidad de que un ejemplo pertenezca a la clase positiva. La predicción final se decide según un umbral (por lo general 0.5)



UNIVERSIDAD EAFIT

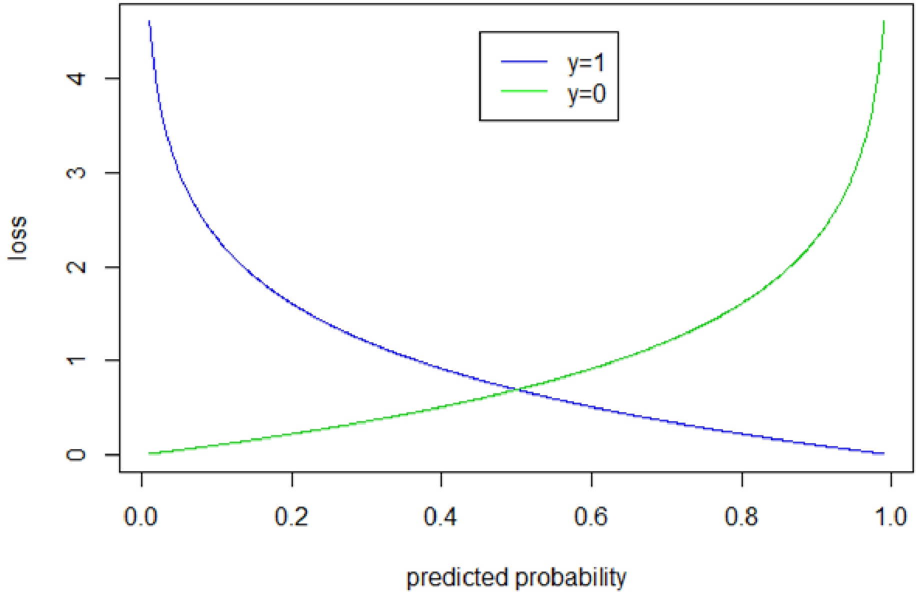
Regresión Logística: Función Objetivo

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

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

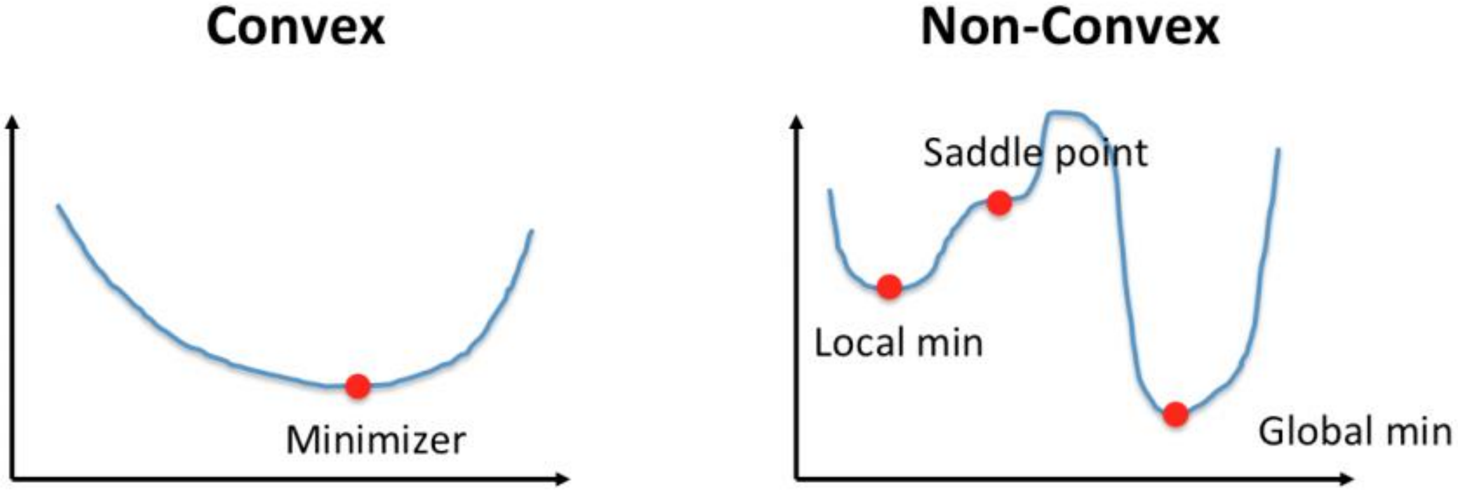
Regresión Logística: Función Objetivo

Esta función objetivo se denomina **pérdida de entropía cruzada (cross-entropy loss)**.



Regresión Logística: Función Objetivo

La pérdida de entropía cruzada se utiliza porque permite que el problema de optimización para entrenar la regresión logística se vuelva **convexo**.



Regresión Logística: Entrenamiento

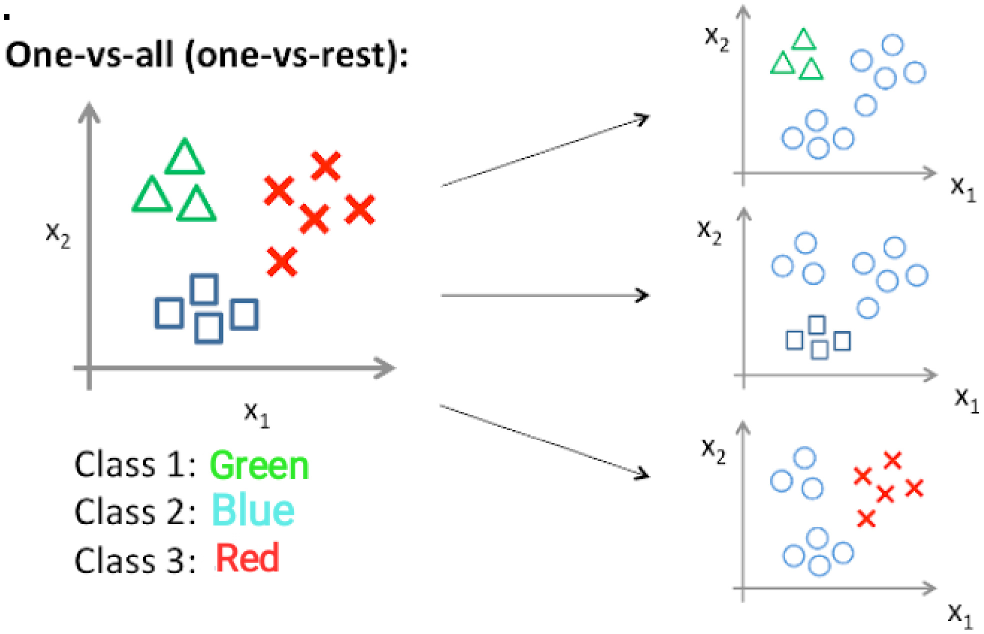
La regresión logística sólo se puede entrenar mediante métodos iterativos. Por lo general, se utiliza el descenso por el gradiente.

$$\mathbf{w} := \mathbf{w} - \alpha \times \nabla Cost_{\mathbf{w}}$$

$$\nabla Cost_{\mathbf{w}} = \frac{1}{m} X^T(\hat{y} - y)$$

Regresión Logística: Multiclase

Para lograr una clasificación multiclase mediante regresión logística se puede aplicar la estrategia **uno contra todos (one-vs-all o one-vs-rest)**.



Regresión Logística: Multiclase

Otra opción para lograr una clasificación multiclase es aplicar una **regresión logística multinomial**, también conocida como **regresión softmax**.

$$h_\theta(x) = \begin{bmatrix} P(y=1|x;\theta) \\ P(y=2|x;\theta) \\ \vdots \\ P(y=K|x;\theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{K} \exp(\theta^{(j)\top} x)} \begin{bmatrix} \exp(\theta^{(1)\top} x) \\ \exp(\theta^{(2)\top} x) \\ \vdots \\ \exp(\theta^{(K)\top} x) \end{bmatrix}$$

Preprocesamiento

Preprocesamiento

Tanto la regresión lineal como la logística funcionan mejor si las entradas varían en el mismo rango de valores. Debido a esto, se recomienda transformar las características para que varíen entre 0 y 1 o entre -1 y 1 o para que tengan media 0 y varianza 1 (estandarización). En Scikit-Learn esto se puede lograr con las siguientes transformaciones:

- sklearn.preprocessing.MinMaxScaler
- sklearn.preprocessing.StandardScaler

Ejemplo de regresión lineal y logística

GRACIAS

UNIVERSIDAD EAFIT