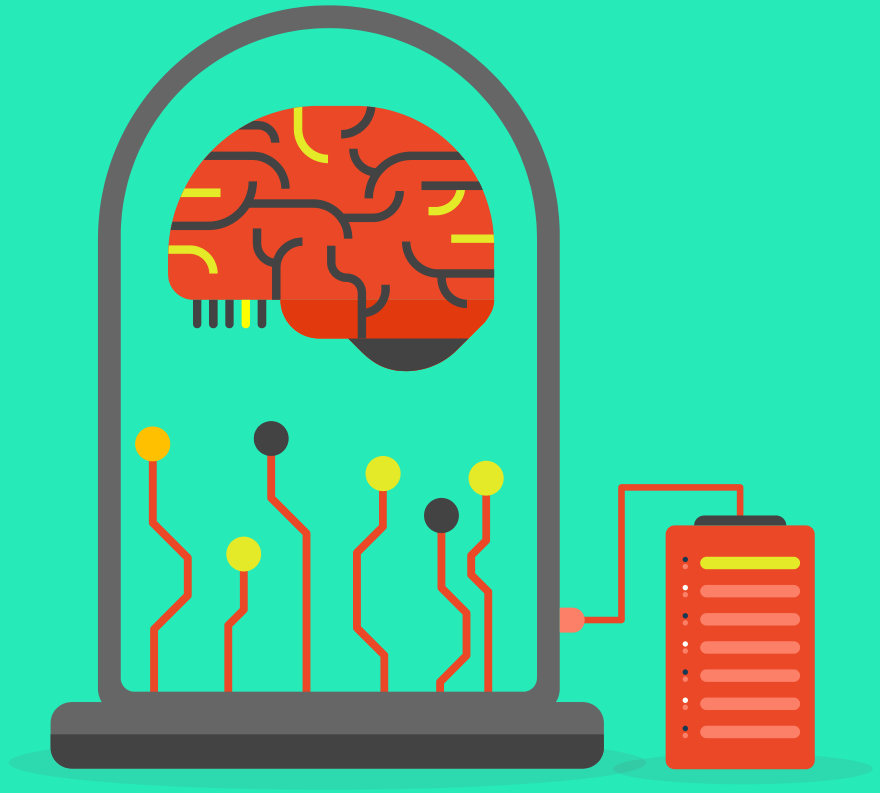
11. Regresión Logística



Estimación de Probabilidades

Supongamos que deseamos construir un clasificador capaz de determinar si un dato dado pertenece o no pertenece a una clase. La **Regresión Logística** es un tipo de modelo de Machine Learning que permite estimar la probabilidad de que un registro de un conjunto de datos pertenece a una clase en particular. Si para dicho dato, la probabilidad estimada es superior al 50%, entonces el registro sí pertenece (se etiqueta como 1) de lo contrario, no pertenece (se etiqueta como 0). Se cuenta entonces con un clasificador binario.

Recordando la hipótesis de la Regresión Lineal:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Que en su forma vectorizada se puede escribir como:

$$h_\theta(x) = x^T \theta$$

Estimación de Probabilidades

La Regresión Logística plantea como salida del modelo, no esta última cantidad directamente, sino la *logística* de la misma:

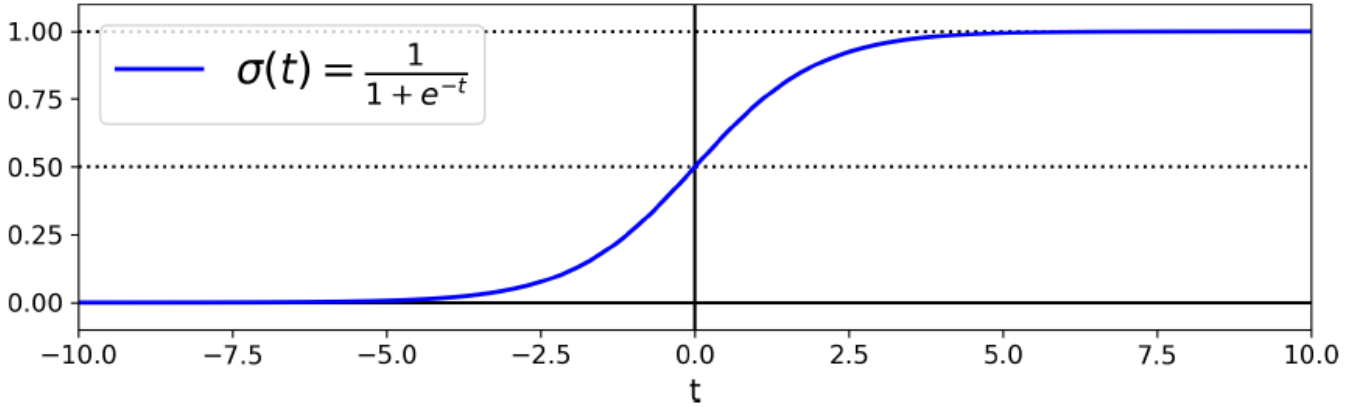
$$\hat{p} = h_\theta(x) = \sigma(x^T \theta)$$

En donde:

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$

Estimación de Probabilidades

La forma de esta función es la siguiente:



Y para la Regresión Logística, las predicciones estarán dadas por la regla:

$$\hat{y} = \begin{cases} 0 \; si \; \hat{p} < 0.5 \\ 1 \; si \; \hat{p} \geq 0.5 \end{cases}$$

Función de Costo

En la Regresión Logística, el objetivo del entrenamiento es lograr que los parámetros θ sean tales que el modelo estime probabilidades altas para las clases positivas ($y=1$) y probabilidades bajas para las clases negativas ($y=0$). La función de costo que garantiza esta condición se llama *log loss*, y tiene la forma:

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}\log\left(\hat{p}^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-\hat{p}^{(i)}\right)\right]$$

Esta función de costo no tiene solución analítica cerrada como la ecuación normal, pero sí es continua y diferenciable por lo que se puede minimizar usando el Método del Descenso del Gradiente:

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

Caso Multiclases

Lo anterior aplica para el caso binario. Sin embargo, el modelo puede ser generalizado para clasificaciones multiclases directamente. Para ello, se utiliza un método conocido como *Regresión Softmax,* en donde se calcula la probabilidad de que cada registro pertenezca a cada clase distinta, usando una función conocida como *Softmax*.

Esto, en conjunto con la utilización de una función de costo llamada *Entropía Cruzada* (*Cross Entropy*), garantiza la obtención de un resultado de predicción en donde se obtiene la probabilidad de que un registro pertenezca a cada una de las clases entrenadas.

La matemática de este resultado se escapa del alcance de este curso, pero sí veremos ejemplos prácticos de cómo implementar ambos tipos de regresiones logísticas.

Décimo Primer Notebook Práctico

Veamos el cómo se implementa la regresión logística para un conjunto de datos conocido, y cómo se evalúa su error en función de las métricas estudiadas en la clase pasada.

Décimo Primer Notebook Práctico