Apuntes de Matemáticas

Regresión lineal

El modelo predictivo en una regresión lineal se asume que la variable a predecir Y es una función lineal de la o las variables de entrada X. Es decir que,

$$y_e = \alpha + \beta x$$

En la ecuación aparecen dos parámetros α y β que se calcularán para minimizar el error entre las variables de entrada y de salida, de tal manera que podamos encontrar cualquier valor de Y.

Esto, matemáticamente se traduce en la siguiente expresión:

$$e_i = (y_i - Y_e(x_i))$$

Entonces, el objetivo es minimizar la suma de los errores sobre todos los puntos del data set. Todos estos puntos son los valores de x e y de 1 a n:

$$X = \{(x_i, y_i)\}_{i=1}^{n}$$

Para reducir el error, debemos considerar que este puede ser tanto positivo como negativo, por tanto, la expresión del error la debemos calcular como cuadrado:

$$min \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - Y_e(x_i))^2 = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Utilizando el cálculo diferencial se puede demostrar que los parámetros que minimizan la ecuación anterior vienen dados por:

$$\beta = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

Es decir,

$$\beta = \frac{Cov(x,y)}{Var(x)}$$

Y, por tanto, al despejar

$$\alpha = \overline{y} - \beta \overline{x}$$

La componente de error

En la realidad, un modelo lineal no podrá predecir con exactitud todos los valores, por tanto se considerará que el modelo **siempre tiene una componente de error, que responderá a una variable cualquiera con (necesariamente) una distribución normal.**

$$y = \alpha + \beta x + \varepsilon$$

En el caso de que la variable residual ε no responda a una distribución normal, deberemos desechar nuestro modelo lineal, pues en el fenómeno se estará produciendo una distribución (logarítmica, exponencial, cuadrática, etc.) que nuestro modelo no está siendo capaz de representar.