Medidas de distancia en datos

LECCIÓN 2

La distancia es una forma de cuantificar la diferencia entre dos o mas datos. Si bien el término distancia puede suponer que solo se miden distancias típicas como por ejemplo, qué tanto se aleja un punto de otro o un punto del origen de coordenadas, sin embargo, existen diferentes medidas de distancias en función de las diferencias que se quieran cuantificar. Por ejemplo, para saber la diferencia que hay entre dos palabras se requiere una medida distancia diferente a la diferencia entre dos puntos. También cambia si se necesita la distancia entre dos clusters de datos. Al estimar qué tan lejos está un punto de otro bajo un criterio (la medida de distancia) es posible saber responder preguntas importantes en ciencia de datos como son: ¿Se aproxima mi modelo a los datos? ¿Los parámetros reducen el error de la estimación de los datos? ¿El costo o diferencia entre los datos de entrenamiento y las estimaciones del modelo se alejan o se acercan a la realidad? Lo que, en esencia, permite entrenar modelos y optimizar modelos de aprendizaje de máquina.

Algunas de las medidas típicas de distancia son:

Distancia euclidiana

Es la distancia entre dos puntos de un espacio euclidiano. Para dos puntos (x1, y1) y (x2, y2) la distancia se calcula como:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Distancia Manhattan

Esta distancia se calcula teniendo en cuenta las diferencias absolutas entre las coordenadas de dos puntos, es decir, se inspira en las ciudades que están construidas por cuadras y para ir de un punto a otro se debe recorrer externamente cada cuadra en lugar de pasar en medio de ellas (lo que se calcularía con la distancia euclidiana).

$$d = |x_2 - x_1| + |y_2 - y_1|$$

Distancia de Hamming

Es empleada para calcular la diferencia entre dos secuencias de igual longitud. Se cuenta el número de posiciones en las cuales los símbolos correspondientes son diferentes. Por ejemplo, para las secuencias x y y, la distancia d se calcula como:

$$d = \sum_{\{i=1\}}^{n} \delta(x_i, y_i)$$

En donde la función delta, para las secuencias xi y yi devuelve cero si son iguales en ese punto o devuelve un uno si son diferentes las secuencias en el punto i. Se puede emplear para calcular qué tan diferente es una palabra de otra en un texto (la palabra es una secuencia) y sugerir una corrección ortográfica acertada, con base en las distancias.

Distancia de Chebyshev

Esta distancia se define como la mayor diferencia entre las coordenadas de los puntos en cada dimensión. Para dos puntos (x1, y1), (x2, y2) se calcula como

$$d = \left(\left|x_1 - x_2\right|, \left|y_1 - y_2\right|\right)$$

Distancia de Canberra

es una medida ponderada que se emplea en análisis multivariado. Permite estimar diferencias cuando las variables tienen diferentes escalas. Para dos vectores X y Y, la distancia se calcula como:

$$d = \sum_{\{i=1\}}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|}$$

Distancia Coseno

Mide el coseno del ángulo entre dos vectores en un espacio multidimensional. Se emplea para calcular la diferencia entre conceptos en modelos de análisis de textos. Por ejemplo, con la distancia coseno es posible calcular las diferencias que hay entre ideas como hombre – mujer, y esta distancia debe ser similar a la que existe entre rey – reina. Con esto, los modelos de lenguaje tienen una codificación que ayuda al entendimiento de las palabras. Se calcula como:

$$d = \frac{xy}{||x||^*||y||}$$

Distancia de Jaccard:

Se emplea para medir qué tan similares son dos conjuntos, generalmente binarios. Es la relación entre el número de elementos comunes entre dos conjuntos y el número total de elementos en los conjuntos. Permite establecer la similaridad y se calcula como:

$$d = 1 - \frac{|A \cap B|}{|A \cup B|}$$

La distancia de Jaccard se puede emplear en casos como la detección de objetos a partir de imágenes, en donde los conjuntos A y B representan los píxeles de un objeto. La distancia medirá qué tan bien el objeto fue detectado.