Distancia de Mahalanobis

Alejandro Jimenez Garzon

February 23, 2023

Introducción

Muchas veces los problemas de clasificación y agrupamiento tienden a definirse a partir de la distancia que tengan dos "vectores" de información en sus respectivas dimensiones. Con la métrica euclidiana se tiene una diferencia directa de cada posible variable u dimensión. Sin embargo esto solo funciona si las dimensiones de cada vector tienen el mismo peso, lo que en la vida real muchas veces no ocurre. En muchos casos las variables presentan correlación y para este tipo de problemas de análisis multivariado, se suele usar una métrica propuesta por el profesor P. C. Mahalanobis en 1936, conocida como la Distancia Mahalanobis. [1]

Suponga que tiene dos grupos con características propias , G_1 y G_2. cada entidad de los grupos puede ser vista como un vector de N características o p-dimensiones. Que se denomina $\mathbf{X}$. Este vector puede ser comparado con un *vector promedio* que contendría los promedios de las variables contenidas en los dos grupos. Por lo que el vector X tendría la misma variación del promedio con respecto a cualquier grupo con el que este relacionado. En otras palabras mira la desviación al promedio total de cada grupo y lo compara con el mismo desvió de un vector de muestra, Si el vector de muestra tiene la misma variación del promedio que uno de los grupos. Se puede considerar del mismo. [2]

Esto puede ser considerado como la versión multivariada de la estandarización regular. $z = \frac{x-\mu}{\sigma}$. la estandarización recae sobre el teorema del valor central, pues es una escala que plantea como las variables x de una población tienden a un valor promedio. En el caso de la distancia de Mahalanobis podríamos ver como los valores se mueven de los promedios de la población y así básicamente determinaríamos que tan lejos esta una entidad de la distribución normal de cada grupo. [2]

Al tomar en cuenta la correlación que se presenta entre las variables se determina el peso de cada una. Pues si las variables no presentaran correlación se podría ver la matriz de covarianza como una matriz identidad donde el escalamiento seria básicamente el cuadrado de la distancia euclidiana. Con la inversa de la matriz de covarianza se tiene una forma de de permitir escalar variables que presenten correlación por lo que las estandariza. [2] Con lo anterior planteado la ecuación para el cuadrado de la distancia de mahalanobis esta dado por la siguiente expresión. [1]

$$D^2 = (x - \mu)^T \cdot C^{-1} \cdot (x - m); \tag{1}$$

De la ecuación 1 se tiene x como el vector a comparar con el vector promedio μ multiplicado por la inversa de la matriz de covarianza C^1. Concluyedo que esta métrica puede presentarse como la distancia a la que esta un punto de una distribución de la población. [1]

Aplicaciones

Esta métrica es muy buena en el agrupamiento de datos para encontrar outliers que no pertenezcan a ningún grupo, tambien en la clasificación usar esta métrica mejora la precisión y callback de esta ultima, muchas veces como es el caso de una investigación que clasificaba los tumores entre malignos y benignos se encontró que al usar esta métrica en la matriz de confusión se redujeron los casos de falsos negativos a cero. Es una herramienta muy útil en el caso de comparar clases des balanceadas o comparar un vector con todos lo grupos (uno contra todos). [1]

References

[1] Prabhakaran, S. (2022, 1 marzo). Mahalanobis Distance – Understanding the math with examples (python). Machine Learning Plus. https://www.machinelearningplus.com/statistics/mahalanobis-distance/

[2] McLachian, G. F. (1999, junio). Mahalanobis Distance. Indian Academy of Sciences. Recuperado 22 de febrero de 2023, de https://www.ias.ac.in/article/fulltext/reso/004/06/0020-0026