Instituto Tecnológico de Tijuana

Ingeniería en Sistemas Computacionales

Investigación I:

Distancia Euclidiana

Materia: Minería de Datos

Unidad: Unidad III

Facilitador:

José Christian Romero Sánchez

Alumno: Hernández Negrete Juan Carlos

Fecha:

Tijuana Baja California a 01 de Junio del 2021

What is the goal of the Euclidean distance for the K-Nearest Neighbors (K-NN) machine learning model?

K Nearest Neighbors is one of the most basic and essential classification algorithms in Machine Learning. It belongs to the domain of supervised learning and finds intense application in pattern recognition, data mining, and intrusion detection.

The KNN classifier is also an instance-based, non-parametric learning algorithm:

- **Non-parametric:** means that you do not make explicit assumptions about the functional form of the data, avoiding you mis-model the underlying distribution of the data. For example, suppose our data is highly non-Gaussian, but the Machine Learning model we choose assumes a Gaussian form. In this case, our algorithm would make extremely poor predictions.
- **Instance-based learning** means that our algorithm does not explicitly learn a model. Instead, he chooses to memorize the training instances that are later used as "knowledge" for the prediction phase. Specifically, this means that only when a query is made to our database, that is, when we ask it to predict a label with an input, will the algorithm use the training instances to give an answer.

It should be noted that the KNN minimal training phase is performed both at a memory cost, since we must store a potentially huge data set, and a computational cost during the test time, since the classification of a given observation requires a exhaustion of the entire data set.

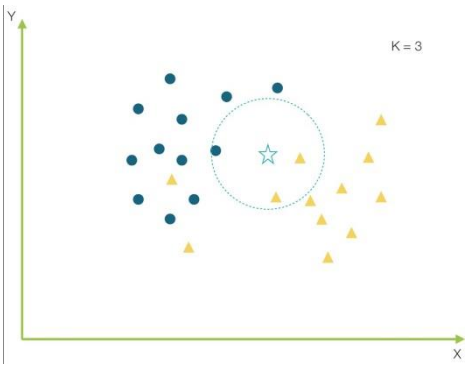

Figure 1 (Example of K-NN Neighbors)

Explain the Euclidean distance equation in your own words

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{p} (x_{ri} - x_{rj})^2}$$

Formula (Euclidean Distance)

For the formula, it is easy to explain, since it is based on the Pythagorean theorem, for this, the distance of the points (P) are in some space of n values, therefore it is considered B-dimensional, for this it is needs the values of X and Y, then these data are taken to be able to apply them in the formula, K being the number of points to consider, once its value is defined, a circle is created relative to the points and the distance there is measured between them, like K, then it can be defined that these values are the closest neighbors, in order to be able to group them according to the distances between them.

References

- Blog de Machine Learning. (2020). K Vecinos más Cercanos – Teoría. 02 de Junio del 2021, de AprendeIA Sitio web: https://aprendeia.com/k-vecinos-mas-cercanos-teoria-machine-learning/#:~:text=K%20vecinos%20m%C3%A1s%20cercanos%20es,y%20la%20detecci%C3%B3n%20de%20intrusos.
- Merkle Inc.. (Septiembre 01 del 2020). El algoritmo K-NN y su importancia en el modelado de datos. 02 de Junio del 2021, de Merkle Inc. Sitio web: https://www.merkleinc.com/es/es/blog/algoritmo-knn-modelado-datos