K-nearest neighbor

K-nearest neighbor (KNN): is finding the outcome of a new data point only based on K nearest data points in training group.

This *will not learn* anything from training data, it will remember all of them.

> *In my opinion, KNN will be effective to classify or predict the group that each element have the similar characteristics and the new data does not tend to change very much compared to the trained group and its characteristics is like everlasting (some model can be used like: iris classification, animals, furniture stock,..)*

💡 Use it when need to predict outcome of new data, can be used both in *regression* and *classification.*

In classification

Label of a new data point can be deduced from K nearest data points in train group. It can be decided by *major voting* or *by weighting different weights.*

In regression

Outcome of a data point will be *outcome of the nearest known data point (if K=1)* or *average weights of nearest outcome data point*, or *the relationship based on nearest data points and distance to them*.

How it works?

Calculate the distance from the new data point to all N given data point then choose the K min distance. If does not have the effective calculation, the calculation will be very large.

Use Euclid to define the distance between 2 data point in multidimensional space.

Khoảng cách từ một điểm tới từng điểm trong một tập hợp

Khoảng cách Euclid từ một điểm $\mathbf{z}$ tới một điểm $\mathbf{x}_i$ trong tập huấn luyện được định nghĩa bởi $\|\mathbf{z}-\mathbf{x}_i\|_2$. Vì trong cách tính này có một bước phải tính căn bậc hai nên người ta thường tính $\|\mathbf{z}-\mathbf{x}_i\|_2^2$. Nếu việc tính khoảng cách chỉ để phục vụ việc sắp xếp thì ta không cần tính căn bậc hai sau bước này nữa. Để ý rằng

$$\|\mathbf{z}-\mathbf{x}_i\|_2^2 = (\mathbf{z}-\mathbf{x}_i)^T(\mathbf{z}-\mathbf{x}_i) = \|\mathbf{z}\|_2^2 + \|\mathbf{x}_i\|_2^2 - 2\mathbf{x}_i^T\mathbf{z} \tag{9.1}$$

However the more effective way to calculate the distance is using the function ***cdist of scipy.spatial.distance*** or ***pairwise_distances in sklearn.metrics.pairwise.***

9.4.2 Ưu điểm của KNN

1. Độ phức tạp tính toán của quá trình huấn luyện là bằng 0.
2. Việc dự đoán kết quả của dữ liệu mới rất đơn giản (sau khi đã xác định được các điểm lân cận).
3. Không cần giả sử về phân phối của các class.

9.4.3 Nhược điểm của KNN

1. KNN rất nhạy cảm với nhiễu khi K nhỏ.
2. Như đã nói, KNN là một thuật toán mà mọi tính toán đều nằm ở khâu kiểm thử. Trong đó việc tính khoảng cách tới *từng* điểm dữ liệu trong tập huấn luyện tốn rất nhiều thời gian, đặc biệt là với các cơ sở dữ liệu có số chiều lớn và có nhiều điểm dữ liệu. Với K càng lớn thì độ phức tạp cũng sẽ tăng lên. Ngoài ra, việc lưu toàn bộ dữ liệu trong bộ nhớ cũng ảnh hưởng tới hiệu năng của KNN.