MODULE II

Classification Vs Regression

The machine learning algorithms are primarily divided into *classification* and *regression* algorithms.

Regression

- Regression algorithms predict a continuous value based on the input variables.
- The main goal of regression problems is to estimate a mapping function based on the input and output variables.
- If your target variable is a quantity like income, scores, height or weight, or the probability of a binary category (like the probability of rain in particular regions), then you should use the regression model.

Classification

- Classification is a predictive model that approximates a mapping function from input variables to identify discrete output variables, which can be labels or categories.
- The mapping function of classification algorithms is responsible for predicting the label or category of the given input variables.

Nearest Neighbor Search

- It is a form of proximity search
- It is the optimisation problem to find the point which close to the given point
- NN search problem:
    - Given a set S of points in a space M and a query point q element of M.
    - Find the close point in S to q
    - This is also called a post office problem
    - Direct generalisation of this problem is a KNN algorithm we need to find K clos path.
- Applications : Pattern recognition, statistical classification, computer vision.

Distance Measures

- When assessing how similar two data points , we need to calculate some sort of metric to be able to compare them.
- Distance metric will calculate the distance between two data points
- Distance metric types:
    a. Euclidean Distance
    b. Hamming Distance
    c. City Block (Manhattan) Distance
    d. Square Distance

1. **Hamming distance**

- Hamming distance measures the similarity between two string of the same length.
- The hamming distance is the no: of positions at which the corresponding characters are different

Eg: Two strings, European , American

E	U	R	O	P	E	A	N

A	M	E	R	I	C	A	N
1	2	3	4	5	6	7	8

Here, Hamming distance = 6

- If we have a dataset that is composed of boolean features, then the hamming distance is probably the best to compute the similarity between two data points.
- It is easily calculated by counting the number of positions where the two binary strings have a different value
- Simply XOR

2. Euclidean distance

Euclidean distance is the shortest distance between two points in an N dimensional space also known as Euclidean space.

It is used as a common metric to measure the similarity between two data points and used in various fields such as geometry, data mining, deep learning and others.

Consider two points P1 and P2:   P1: (X1, Y1) P2: (X2, Y2)

Then, the euclidean distance between P1 and P2 is given as:

$$\sqrt{(x1 - x2)^2 + (y1 - y2)^2}$$

Euclidean distance in N-D space

In a N dimensional space, a point is represented as (x1, x2, ..., xN).

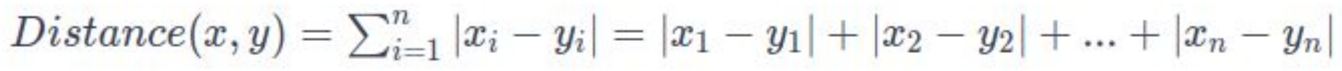
Consider two points P1 and P2: P1: (X1, X2, ..., XN) P2: (Y1, Y2, ..., YN)

Then, the euclidean distance between P1 and P2 is given as:

$$\sqrt{(x1 - y1)^2 + (x2 - y2)^2 + \ldots + (xN - yN)^2}$$

3. City Block Distance (Manhattan)

The manhattan distance also known as L1 distance or city block distance calculate the sum of the absolute values of the difference of the coordinates of two points.



$$Distance(x,y) = \sum_{i=1}^{n} |x_i - y_i| = |x_1 - y_1| + |x_2 - y_2| + ... + |x_n - y_n|$$

The manhattan distance calculates how many squares in a grid we would have to go through to get from point A to point B

The distance always return a positive integer.

4. Square Distance

Here, the distance between two vectors is defined as the maximum of the difference between each element of two.

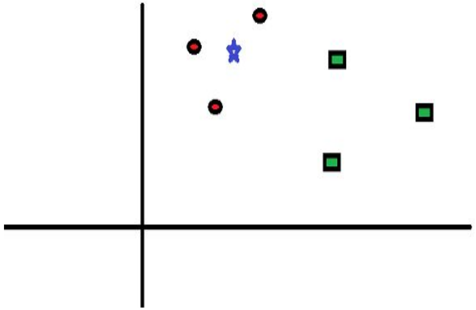
K-Nearest Neighbor

- The K-Nearest Neighbor (KNN) algorithm is a popular machine learning technique used for classification and regression tasks based on supervised learning technique.
- It relies on the idea that similar data points tend to have similar labels or values.
- During the training phase, the KNN algorithm stores the entire training dataset as a reference.
- When making predictions, it calculates the distance between the input data point and all the training examples, using a chosen distance metric such as Euclidean distance.
- Next, the algorithm identifies the K nearest neighbors to the input data point based on their distances.
- In the case of classification, the algorithm assigns the most common class label among the K neighbors as the predicted label for the input data point.
- For regression, it calculates the average or weighted average of the target values of the K neighbors to predict the value for the input data point.
- Non parametric algorithm which means it does not make any assumption on underlying data.
- Also called lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it perform an action on the dataset.

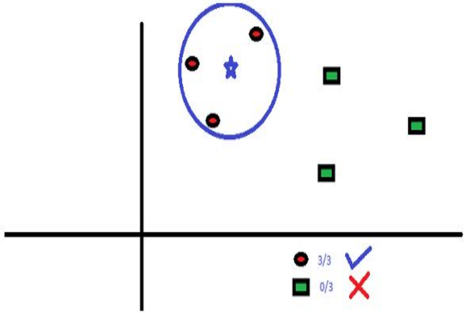
KNN Algorithm

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
    a. Calculate the distance between test data and each row of training dataset. Here we will use Euclidean distance as our distance metric since it’s the most popular method.
    b. Sort the calculated distances in ascending order based on distance values
    c. Get top k rows from the sorted array
    d. Get the most frequent class of these rows
    e. Return the predicted class

- Following is a spread of red circles (RC) and green squares (GS):
- You intend to find out the class of the blue star (BS).
- BS can either be RC or GS and nothing else.
- The "K" in KNN algorithm is the nearest neighbor we wish to take the vote from.
- Let's say K = 3.
- Hence, we will now make a circle with BS as the center just as big as to enclose only three data points on the plane.



- The three closest points to BS are all RC.
- Hence, with a good confidence level, we can say that the BS should belong to the class RC.
- Here, the choice became obvious as all three votes from the closest neighbor went to RC.
- The choice of the parameter K is very crucial in this algorithm.
- Next, we will understand the factors to be considered to conclude the best K.



How Do We Choose the Factor K?

- We run KNN algorithm several times with different values of K and reduce the number of errors .
- Where k=1, the predictions are less stable.
- When the values of k increase , the prediction become more stable due to majority of voting, pusing k so far.
- But it will increase the complexity of algorithm.
- In classification problem, we need majority voting by k into odd number

Advantages

- It is simple to implement
- It is robust to the noisy training data
- More effective when training data is small

Disadvantages

- Always need to determine the values of k which is complex.
- The computational cost is high for calculating distance value between al data points
- Large dataset take longer to process