K-Nearest Neighbors

- K-nearest neighbor (KNN) algorithm is a simple, easy to implement supervised machine learning algorithm that can be used to solve both classification and regression problems.
- This algorithm considers K-nearest neighbors (data points) to predict the class or continuous value for the new data point.
- The KNN is a non-parametric algorithm, which means that it does not make any assumption on the underlying data.
- It is also called **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset at the time of classification, it performs an action on the dataset. The KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Parametric Vs Non-parametric Models

- Parametric model estimates a fixed number of parameters from the data, and they have strong assumptions about the data.
- The data is assumed to be following a specific probability distribution. (Normal or Gaussian) e.g. Naïve Bayes
- Non-Parametric model does not make any assumptions on the type of data distribution.
- k-NN is a non-parametric model which memorizes the data and gives the output for the new observations by comparing the training data.

How does K-Nearest Neighbors work?

- The principle behind K-nearest neighbors is that the nearest neighbors are those data points that have minimum distance in the feature space from the new data point.
- K is the number of such data points that we consider in the implementation of this algorithm.
- For predicting class or continuous value for a new data point, it considers all the data points in the training dataset.
- For classification: A class label is assigned to the majority of K Nearest Neighbors from the training dataset is considered as a predicted class for the new data point.
- For regression: Mean or median of continuous values assigned to K Nearest Neighbors from the training dataset is a predicted continuous value for our new data point.

Nearest Neighbor Classification

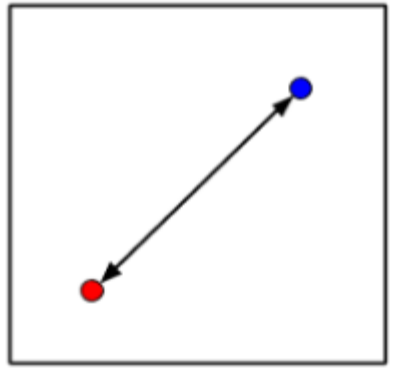
Day	Temperature	Outlook	Humidity	Windy	Play Golf?
07-05	hot	sunny	high	false	no
07-06	hot	sunny	high	true	no
07-07	hot	overcast	high	false	yes
07-09	cool	rain	normal	false	yes
07-10	cool	overcast	normal	true	yes
07-12	mild	sunny	high	false	no
07-14	cool	sunny	normal	false	yes
07-15	mild	rain	normal	false	yes
07-20	mild	sunny	normal	true	yes
07-21	mild	overcast	high	true	yes
07-22	hot	overcast	normal	false	yes
07-23	mild	rain	high	true	no
07-26	cool	rain	normal	true	no
12-30	mild	rain	high	false	yes
tomorrow	mild	sunny	normal	false	yes

K-Nearest Neighbors Steps

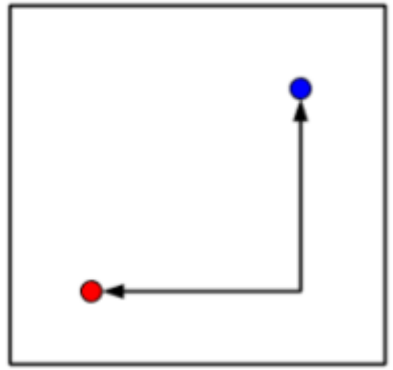
- Step1: Select the number of K neighbors. The most preferred value for K is 5. At very low value of K such as K = 1 or K = 2, can be noisy and lead to the effects of outliers in the model. Large values of K might be good but may find some difficulties.
- Step2: Calculate the Euclidean distance of K number of neighbors.
- Step3: Take the K nearest neighbor as per the calculated Euclidean distance.
- Step4: Among these K neighbors, count the number of the data points in each category.
- Step5: Assign the new data points to that category for which the number of neighbor is maximum.
- Step6: The model is ready.

Distance Metrics

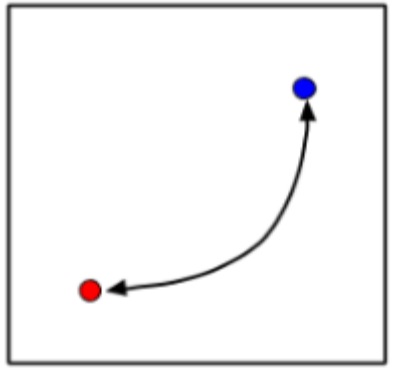
Euclidean



Manhattan



Minkowski



Minkowski distance: a generalization

$$d(i,j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q} \quad (q > 0)$$

If q = 2, d is Euclidean distance

If q = 1, d is Manhattan distance

$$Euclidean = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

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

K-NN Example

- Winterfell Org is a HR consultancy group, and you work as an analyst in that organization. Your job is to identify whether a candidate should be hired or not for the job, based on the interview score and the exam rank. For the new observation, you need to use Euclidean metric to find the distance and consider k=3.

Interview Score	Exam Rank	Type
70	70	Not hired
70	40	Hired
30	40	Not hired
10	40	Not hired
30	70	??

K-NN Example

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

Interview Score	Exam Rank	Euclidean value	Type
70	70	$\sqrt{(70-30)^2 + (70-70)^2}$	Not Hired
70	40	$\sqrt{(70-30)^2 + (40-70)^2}$	Hired
30	40	$\sqrt{(30-30)^2 + (40-70)^2}$	Not Hired
10	40	$\sqrt{(10-30)^2 + (40-70)^2}$	Not Hired
30	70	-	??

Interview Score	Exam Rank	Euclidean value	Type
70	70	40	Not Hired
70	40	50	Hired
30	40	30	Not Hired
10	40	36.05	Not Hired
30	70	-	??

Interview Score	Exam Rank	Euclidean value	Type
30	40	30	Not Hired
10	40	36.05	Not Hired
70	70	40	Not Hired
70	40	50	Hired
30	70	-	Not Hired

Advantages & Disadvantages of K-Nearest Neighbors

- **Advantages**
  - One of the simplest supervised learning method
  - It requires no training time.
  - It can be used for non-linear data

- **Disadvantages**
  - It can become complex with the datasets of higher dimensions.
  - It assumes every feature to have equal importance.
  - It is sensitive to outliers as it finds the distance based on numeric value.

Applications of K-Nearest Neighbors

- **Sports Analytics**
- **Credit Ratings**
- **Healthcare Analytics**

K-NN Regression Problem

Samaritan.io is a leading data consultancy organization in India. Mr. Shah has approached Samaritan.io with the data of 50 startups in India. Mr. Shah is planning to build a startup company and wants to know about the spending they can make on various parameters and the location where they can start the organization. You, being the lead data scientist, is assigned with the task of helping Mr. Shah decide.

R&D Spend	Administration	Marketing Spend	City
150000	200000	380000	Gurgaon
180000	250000	420000	Bangalore
120000	190000	390000	Chennai

Thank You !