Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

K-Nearest Neighbors

K-nearest neighbors is a type of supervised learning algorithm, which is used for both regression and classification purpose; but mostly it is used for classification.

Given a dataset with different classes, KNN tries to predict the crorrect class of test data by calculating the distance between the test data and all the training points.

- Then, it selects the K points which are closest to the test data. Once the points are selected, the algorithm calculates the probability (incase of classification) of the test point belonging to the classes of the k-training points and the class with the highest probability is selected.

Memo No. _____

Mo Tu We Th Fr Sa Su

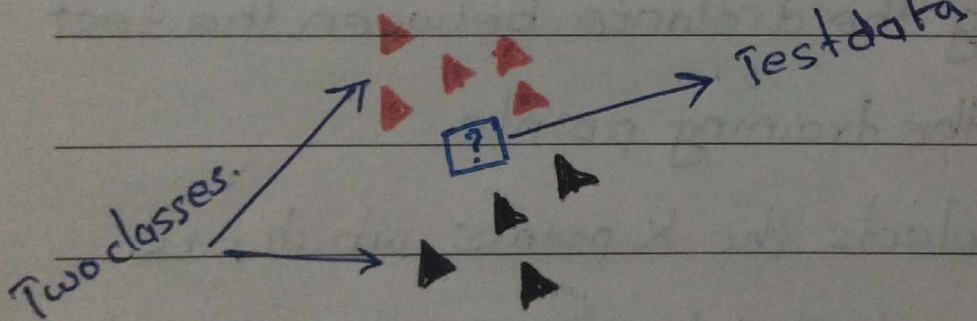
Date / /

In Case of Regression:-

→ The predicted value is the mean of the k-selected training points.

Let's understand this with an illustration:

1) Given a training dataset as given below. We have a new test data that we need to assign to one of the two classes.



2) Now, K-nearest neighbours algorithm calculate the distance between the test data and given training data.

Say, K=5

↓

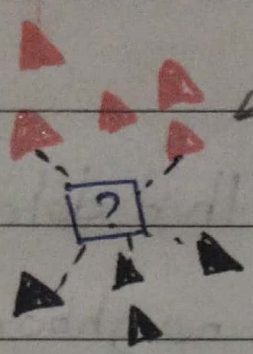
Mean 5 nearest neighbour

other model holds the relationship between dataset.

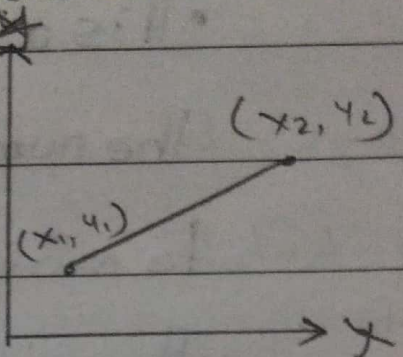
Memo No. _____

Mo Tu We Th Fr Sa Su

Date / /



Here we took 5 nearest point of the test data., the distance is euclidean distance.



$$d = \sqrt{(x_2 - x_1)^2 + (y_1 - y_1)^2}$$

3) Now, calculating the probability of belonginess.

K-nn points = [▲ ▲ ▲ ▲ ▲]

A B

$P(A) = 2/5, \quad P(B) = 3/5.$

Hence, the test data belongs to class 'B'.

Key Point:

K-NN is a lazy learner algorithm.

↓

You will have store entire dataset as a model. So, dataset is itself a model.

↓

If the dataset is too huge, it becomes computationally expensive and slow prediction.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

K-Value:

- It is a hyperparameter that determines the number of nearest neighbors used to make predictions.
- The value of K is typically selected through a process called Cross-validation
- choosing the right value of K is important as a small value of K (eg: K=1) can leads to overfitting where as large value of K (eg: K=n) may result under fitting.

A good heuristic for choosing the value of K is to set it to an odd number if the number of classes is ≤ 2. (To avoid ties.) and squareroot of the number of sample (number of rows) in the dataset

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Types of distances:

a) Euclidean Distance:

The eudidean distance bet^n two points $p(p_1, p_2)$ and $q(q_1, q_2)$ is calculated as.

$$d = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2}$$

for n dimension

$$d = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

b) Hamming distance → It is for categorical

It is used to measure the dissimilarity bet^n instances in dataset. In KNN, the goal is to find the K-nearest neighbors of new instances based on it's feature values. To do this, similarity metric is needed to compare dissimilarity bet^n instances.

ABCD ABCD ABCD ABDC

Memo No. (2)

Date / /

Example of Hamming distance.

A = [0, 1, 0, 1] , B = [1, 0, 1, 0]

Hamming distance beth A & B = 4

A = [0, 1, 0, 1] _ B = [0, 1, 1, 0]

Hamming distance = 2

It can't be directly to categorical data only to binary data

A = [a b c d] & B = [D E F

Hamming distance = 4

$$HamD(x,y) = \sum_{i=1}^{n} 1_{x_i = y_i}$$

c) Manhattan Distance

→ Also known as L1 distance.

In 2 dimensional.

The manhattan distance between (x_1, y_1) and (x_2, y_2) is: $d = |x_1 - x_2| + |y_1 - y_2|$

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

PROS & CONS OF KNN ALGORITHM:

Pros:

- It can be used for both regression and classification problem
- no need to create mode.
- doesn't make any assumption for the distribution.

Cons:

- finding the optimum value of k.
- takes a lot of time to compute distance betⁿ each test sample & all training samples.
- Since, the model is not saved beforehand in this algorithm (lazy learner), so every time one predicts a test value, it follow the same steps again & again.
- Since we have to store the whole training set for every test set, so requires alot of space.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Different ways to perform K-NN:

The above approach is known as 'Brute force K-NN'. This is computationally very expensive.

↓

So, there are other algorithm which are less expensive

↓

aim: to reduce the time during test period by preprocessing the training data in such a way that the test data can be easily classified in the appropriate cluster.

Two famous Algorithm:

ⓐ K-Dimensional Tree (kd-tree):

→ arranged in binary tree structure.

→ while test data is provided, it would give out the result by traversing through the tree, which takes less time than brute force search.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Example:

Training data ⇒ {(1, 2), (2, 3), (2, 4), (3, 6), (4, 2), (5, 7), (6, 8), (7, 5), (8, 5), (9, 1), (9, 3)}

Here $k = 2$.

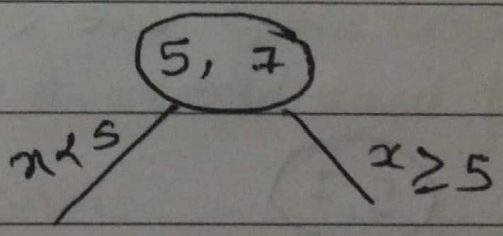
$d = 2$

let's sort our data and choose the median to be the split point:

median

$x \in \{1, 2, 2, 3, 4, \textcircled{5}, 6, 7, 8, 9, 9\}$

1) Now, our first node will be (5, 7).



$S_1 \in \{x < 5\}$	$S_2 \in \{x \geq 5\}$
$S_1 \in \{(1,2), (2,3), (2,4), (3,6), (4,2)\}$	$S_2 \in \{(5,7), (6,8), (7,5), (8,5), (9,1), (9,3)\}$

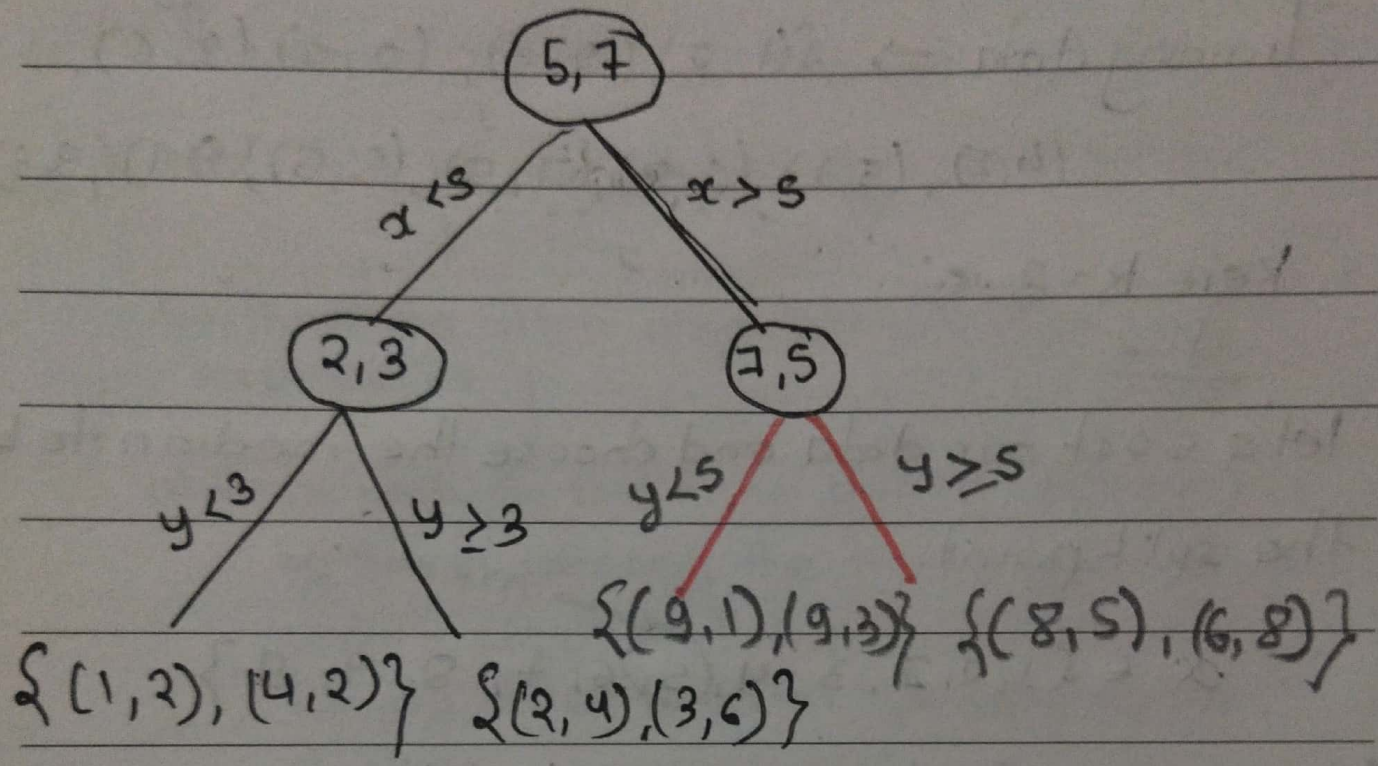
2) let's spit S_1 and S_2 on condition of y.

$y_{S_1} \in \{2, 2, \textcircled{3}, 4, 6\}$ $y_{S_2} \in \{1, 3, \textcircled{5}, 5, 8\}$

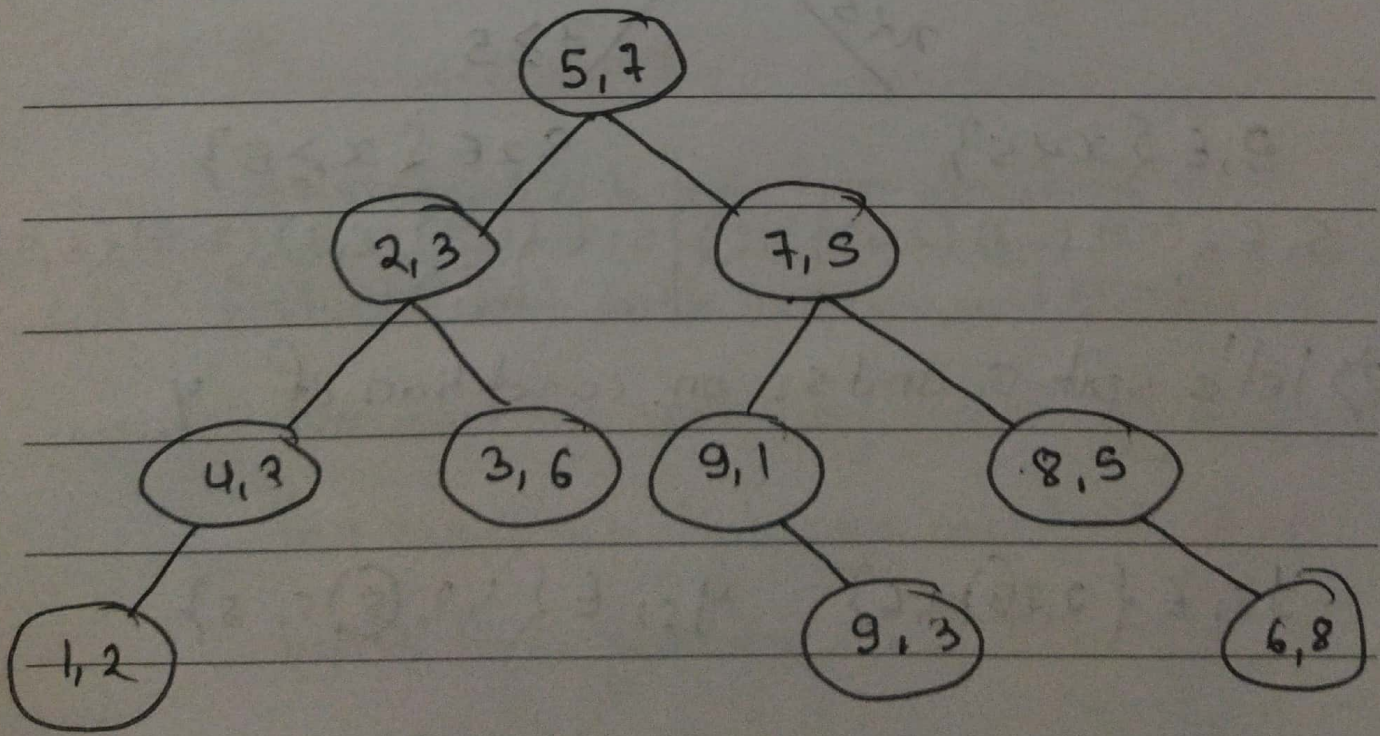
Memo No.

Date / /

Now building tree:



Similarly, in next step use x to split & the final will look like:



Memo No. _____

Mo Tu We Th Fr Sa Su

Date / /

Now, if our test data is (8, 6) then (7,5), (8, 5) (6, 8) would be neighbour.

b) Ball Tree: ← similar to KMeans clustering

- Similar to K-d tree, it is also hierarchical data structure
- Efficient for higher dimension

Steps:

1) Two clusters are create initially.
2) All the datapoints must belong to one of the clusters.
3) One point cannot be in both clusters.
4) Distance of the point is calculated from the centroid of the each cluster. The point closer to the centroid goes into the particular cluster.
5) Each cluster is then divided into sub-clusters again & then the points are

Memo No. _____

Mo Tu We Th Fr Sa Su

Date / /

classified into each cluster on the basis of distance from centroid.

6) This is how the clusters are kept to be divided into certain depth.