Unit - IV

# Ensemble Techniques & Unsupervised Learning.

## 1.) Bagging.

* Bagging is also called bootstrap aggregating. Bagging and boosting are meta-algorithms that pool decisions from multiple classifiers.

* The ensemble learning method that is commonly used to reduce variance within a noisy dataset.

* The meta-algorithm, which is a special case of the model averaging was originally designed for classification and is usually applied to decision tree models

* Ensemble classifiers such as bagging, boosting and model averaging are known to have improved accuracy and robustness over a single model.

* Each bootstrap sample will on average contain 63.2% of the unique training examples.

* It combines with m resulting models using simple majority vote.

* The base learner is trained on what is often called a bootstrap replicate.

* It decreases error by decreasing the variance in the results due to unstable learners

**Pseudo code :**

i) Given training data $(x_1, y_1) \ldots (x_m, y_m)$

ii) For $t = 1, \ldots T$

a) Form bootstrap replicate dataset $S_t$ by selecting $m$ random examples from the training set with replacement.

b.) Let $h_t$ be the result of training base learning algorithm on $S_t$.

iii) Output combined classifier

$$H(x) = \text{majority}(h_1(x), \ldots h_T(x)).$$

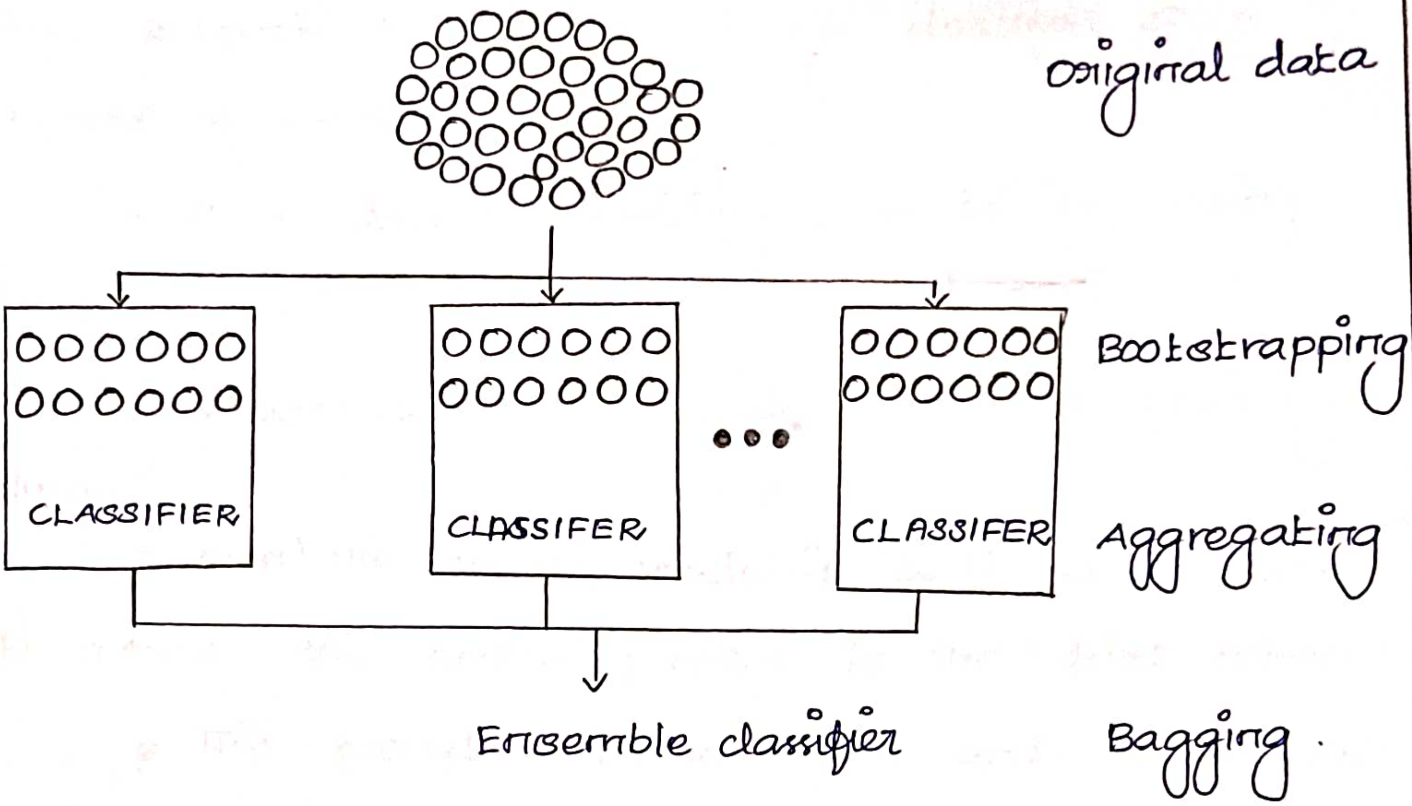
**Bagging steps :**

**Step 1 :** Multiple subsets are created from the original data set with equal tuples, selecting observations with replacement.

**Step 2 :** A base model is created on each of these subsets.

**Step 3 :** Each model is learned in parallel with each training set and independent of each other.

Step 4 : The final predictions are determined by combining the predictions from all the models.




## Advantages

* Reduces Over-fitting of the model
* Handles higher dimensionality data very well.
* Maintains accuracy for missing data.

## Disadvantages

Since final predictions is based on the mean predictions from subset trees, it wont give precise values for the classification and regression model.

# Boosting

* Boosting is an ensemble modeling technique that attempts to build a strong classifier from the number of weak classifiers.

* It is done by building a model by using weak models in series.

* Firstly, a model is built from the training data.

* Then the second model is built which tries to correct the errors present in the first model

* This procedure is continued and models are added until either the complete training data set is predicted correctly.

Types of boosting algorithms.

* Gradient boosting

* XGBoost

* Adaboost

* Catboost.

Adaboost.

* Adaboost, short for "Adaptive boosting" is a machine learning meta.

* It can be used to learn weak classifier and final classification based on weighted vote of weak classifiers.

* It is linear classifier with all its desirable properties.

* All weights are set equally, but each round the weights of incorrectly classified.

Advantages of Adaboost.

* very simple to implement.
* Fairly good generalization
* The prior error need not be known ahead of time

Disadvantages of Adaboost.

* Suboptimal solution
* can over fit in presence of noise.

## Training of boosting model.

i) Initialise the dataset and assign equal weight to each of the data point.

ii) provide this as input to the model and identify the wrongly classified data points.

iii) Increase the weight of the wrongly classified data points

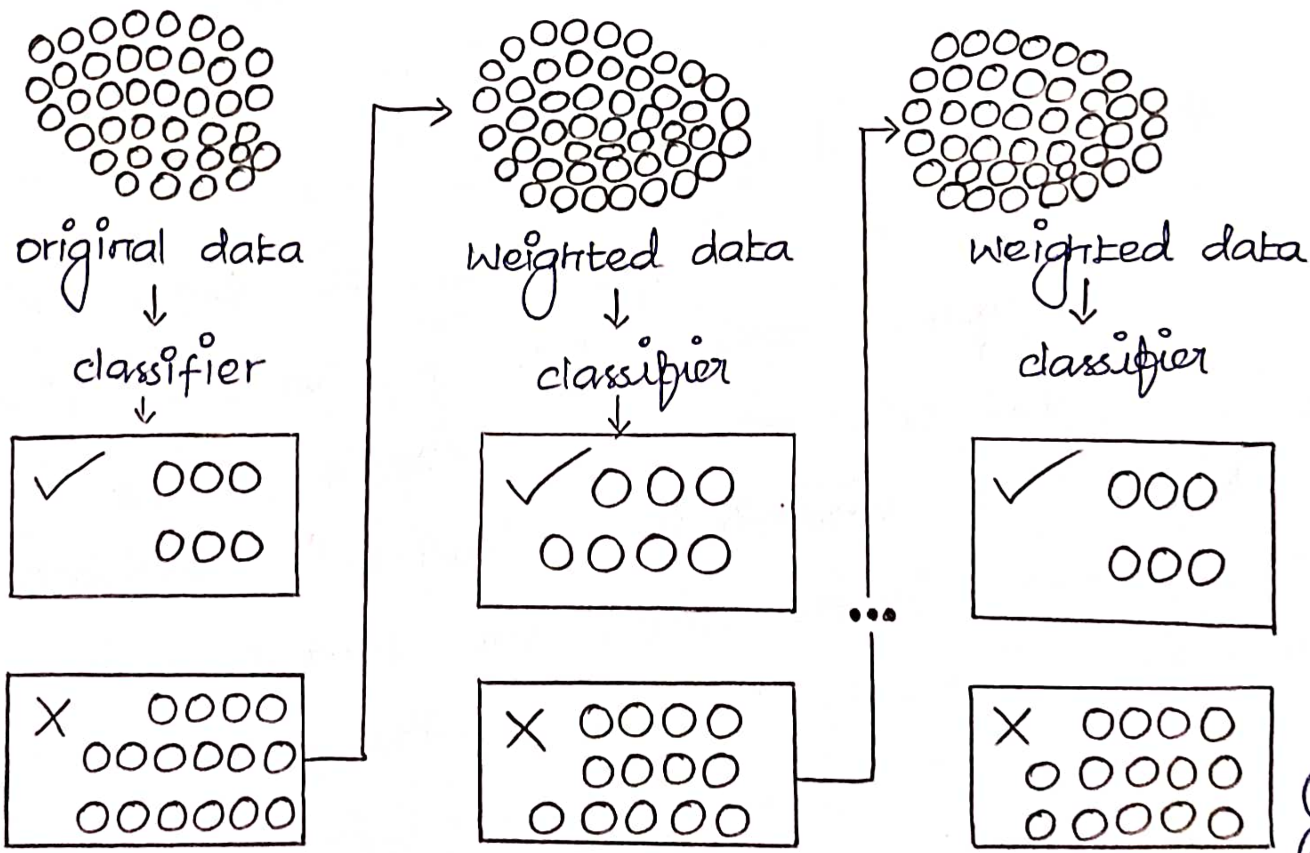
iv) if (got required results)

Goto step 5

else

Goto step 2

v) End




6

8

Advantages

i) Supports different loss function

ii) Works well with interactions.

Disadvantages

i) Prone to over-fitting

ii) Requires careful tuning of different hyper-parameters.

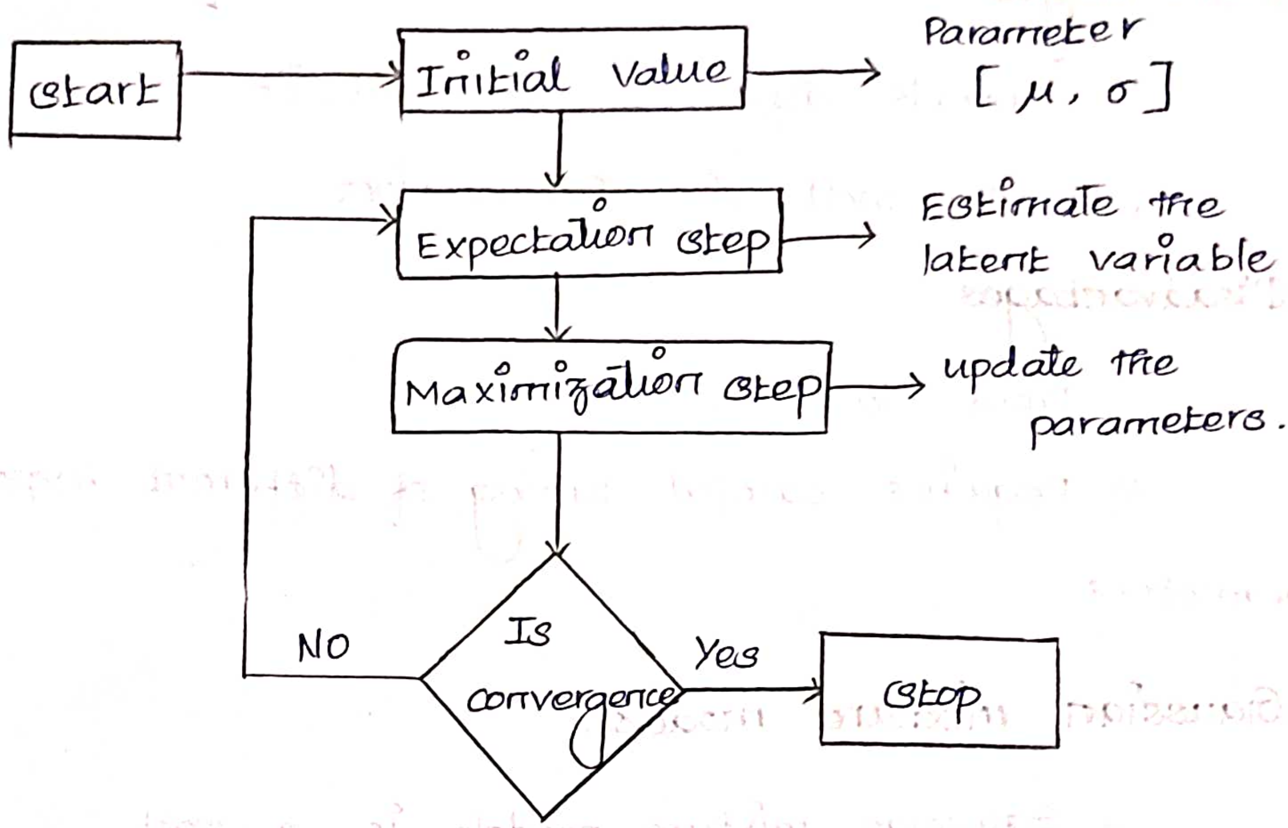
2) Gaussian mixture models

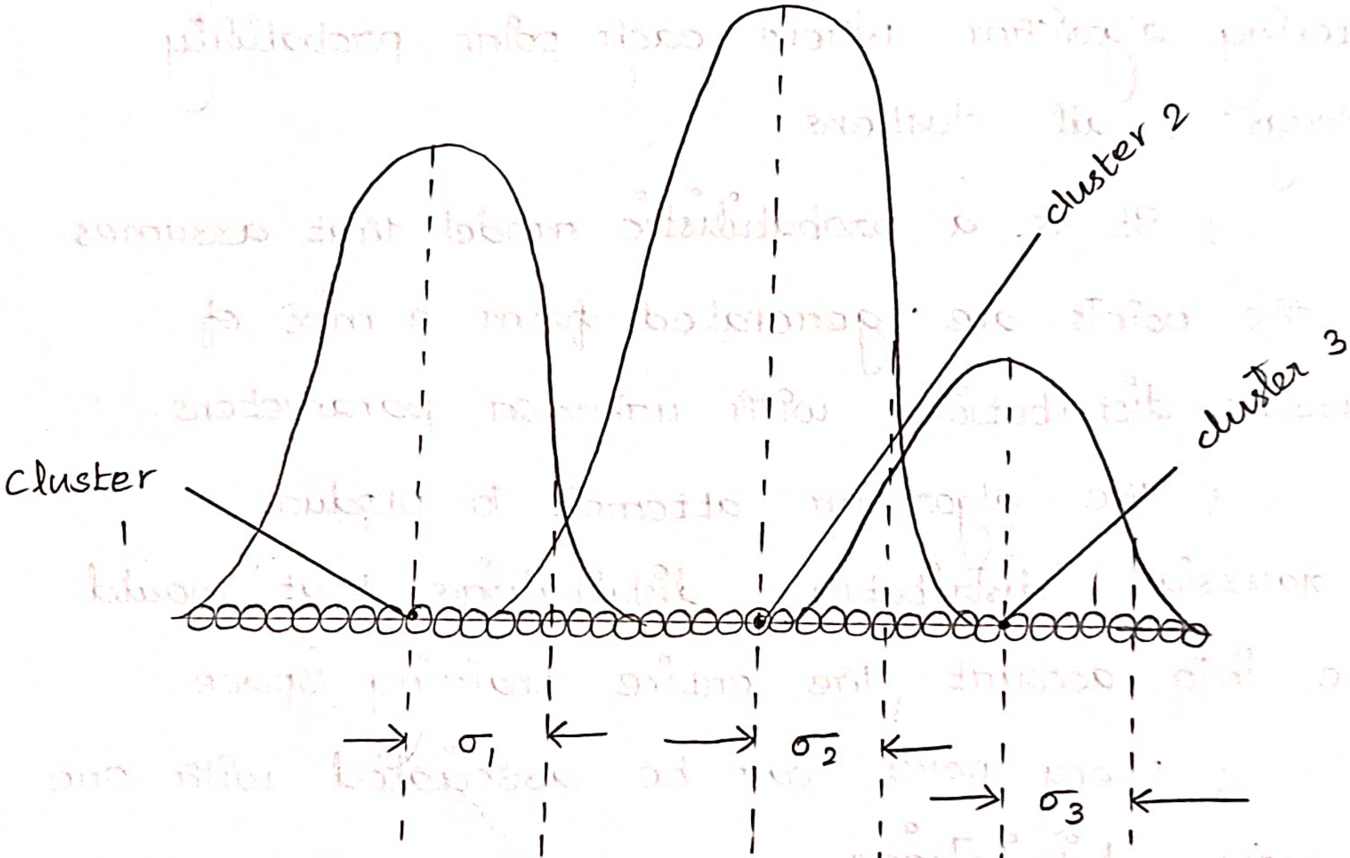
⇒ Gaussian mixture models is a soft clustering algorithm. Where each point probability "belongs" to all clusters

⇒ It is a probabilistic model that assumes all the points are generated from a mix of Gaussian distributions with unknown parameters.

⇒ The algorithm attempts to produce k-Gaussian disbubutions distributions that would take into account the entire training space

⇒ Every point can be associated with one or more distributions.




It is used to determine the probability of each point belongs to a given cluster.

⇒ Gaussian mixture models have a variety of real - world applications

a) Used for signal processing

b) Used for customer churn analysis

c) Used for Language identification

d) Used in video game industry

e) Genre classification of songs.

Ex :

In modeling human height data, height is typically modeled as a normal distribution for each gender with a mean of approximately 5'10" for males and 5'5" for females. here given Only the height data and not the gender assignments for each data point, the distribution of all heights would follow the sum of two scaled and shifted normal distributions. A model making this assumption is an example of a Gaussian mixture model.

## Expectation - Maximization Algorithm

⇒ The Expectation - Maximization algorithm is an iterative way to find maximum likelihood estimates for model parameters when the data is incomplete.

⇒ Expectation - Maximization chooses some random values for the missing data points and estimates a new set of data.

⇒ These new values are then recursively used to estimate a better first date, by filling up missing points, until the values get fixed.

⇒ There are two most important steps that are iteratively performed.

### Estimation step.

⇒ we first initialize our model parameters like ther mean covariance matrix and mixing coefficients.

⇒ calculate the posterior probabilities of data points belonging to each centroid.

⇒ These probabilities are often represented by the latent variables yk.

Maximization step

⇒ we update the parameters using the estimated latent variable yk.

⇒ Let update the cluster part and covariance matrix then update the mixing coefficients.

3. K - means and KNN algorithm.

⇒ k - means clustering is heuristic method.

⇒ Each cluster is represented by the center of the cluster.

'k' stands for number of clusters.

⇒ The number of components of the population equal to the final required number of clusters.

⇒ The final required number of clusters is chosen such that the points are mutually farthest apart.

⇒ Everytime a component is added to the cluster. The centroid's position is recalculated.

⇒ This continues until all the components are grouped into the final required number of clusters

⇒ k - means algorithm consists of four steps.

i) Select initial centroids at random.

ii) Assign each object to the cluster with the nearest centroid.

iii) Compute each centroid as the mean of the objects assigned to it.

iv) Repeat previous 2 steps until no change.

$$W(c) = \frac{1}{2} \sum_{k=1}^{k} \sum_{c(i)=k} \sum_{c(j)=k} \| x_i - x_j \|^2$$

$$= \sum_{k=1}^{k} N_k \sum_{c(i)=k}^{k} \| x_i - m_k \|^2.$$

Where

$m_k$ is the mean vector of the $k^{th}$ cluster

$N_k$ is the number of observations in $k^{th}$ cluster.

properties

⇒ There is always atleast one item in each cluster.

⇒ The clusters are non-hierarchical and they do not overlap.

⇒ Every member of a cluster is closer to its cluster.

Advantages

* Efficient in computation.
* Easy to implement.

Disadvantages

* Applicable only when mean is defined.
* Trouble with noisy data and outliers.

KNN algorithm.

⇒ K - Nearest neighbour is one of the only Machine learning algorithms based totally on supervised learning approach.

⇒ K-NN set of rules can be used for regression as well as for classification.

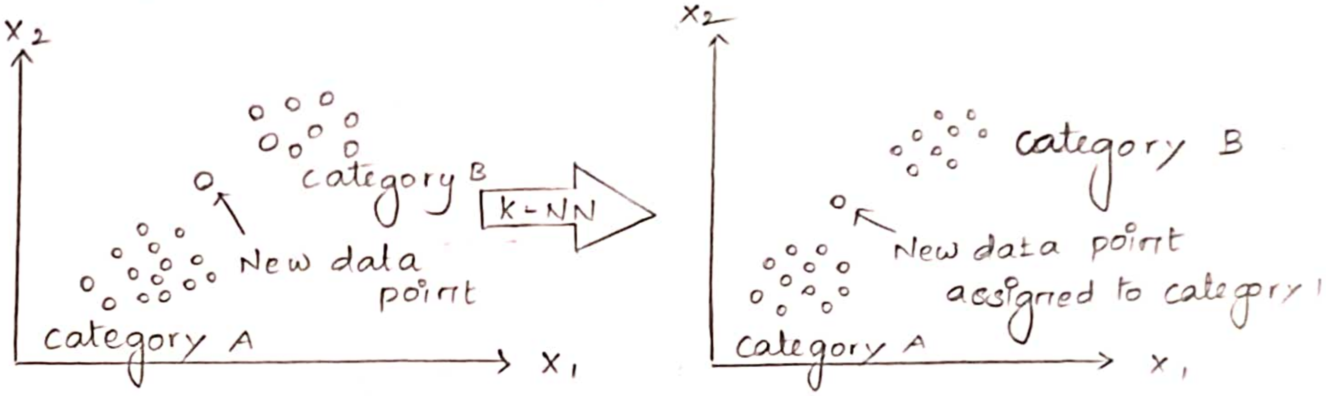
⇒ K-NN is a non-parametric algorithm, because of this it does no longer makes any assumption on underlying data.

Ex :

We've an picture of a creature that looks much like cat and dog but we want to know both it is a cat or dog. So for this identity. we are able to use the KNN algorithm.

## Why do we need KNN ?

There are two categories . categories A and categories B and we've a brand new statistics point $x_1$ . so that fact point will lie within of these classes . to solve this problem we need K - NN . set of rules .




## How does KNN work

step 1 : Select the wide variety k of the acquaintances

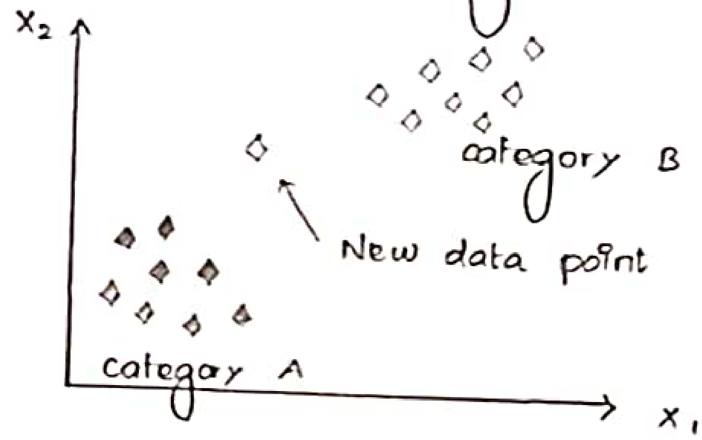
step 2 : Calculate the Euclidean distance of K variety of friends .

step 3 : Take the k - nearest neighbours as according to the calculated Euclidean distance

step 4 : Among these ok pals , count number the number of the data points in each class .

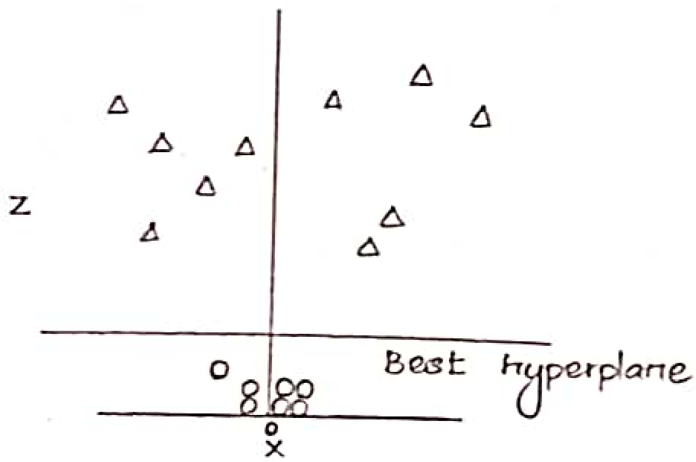
Step 5 : Assign the brand new record points to that category for which the quantity of the neighbour is maximum

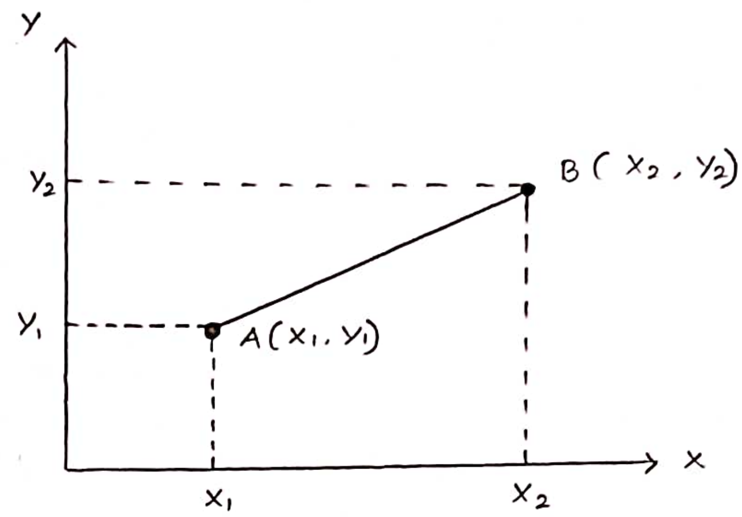
Step 6 : Our model is ready.




* First, we are able to pick the number of friends, so we are able to select the ok = 5.

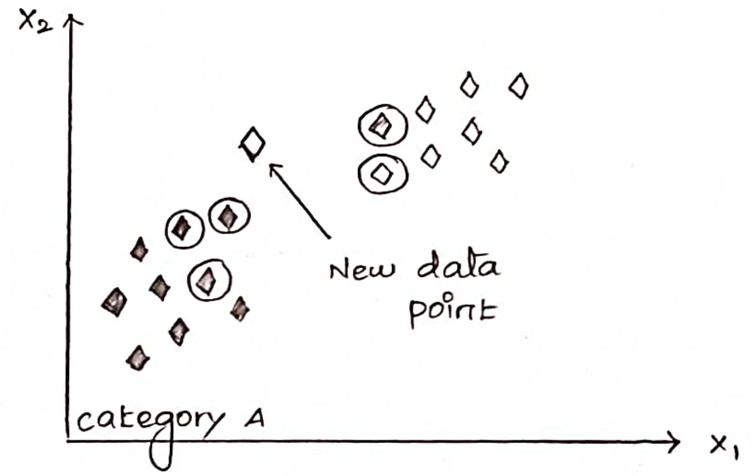
* Next, we will calculate the Euclidean distance between the facts points.






Euclidean distance between $A_1$ & $B_2$ $= \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

* By calulating the Euclidean distance we got the nearest accquaintances.




* Able to see the three nearest acquaintances are from category A. Subsequently this new fact point must belong to category A.