CT-2 DWDM

106118036

1) $X = \{26 \ldots 30, \text{Systems}, 46 \sim 50K\}$

$$P(26 \ldots 30 / \text{juniors}) = \frac{P(\text{Juniors} / 26 \ldots 30) * P(26 \ldots 30)}{P(\text{Juniors})}$$

$$= \frac{1 * 49/165}{113/165}$$

$$= \frac{49}{113}$$

$$P(\text{Systems} / \text{Juniors}) = \frac{P(\text{Juniors} / \text{Systems}) * P(\text{Systems})}{P(\text{Juniors})}$$

$$= \frac{23/34 * 34/165}{113/165}$$

$$= 23/113$$

$$P(46K\text{-}50k / \text{Juniors}) = 23/113$$

$$P(X / \text{Juniors}) \propto 49/113 * 23/113 * 23/113$$

$$\propto \boxed{0.0179}$$

$$P(X / \text{Juniors}) = \frac{49}{113} * \frac{23}{113} * \frac{23}{113} * \frac{113}{165}$$

$$= \boxed{0.0123}$$

$P(\text{Systems} \mid \text{senior}) = \frac{8}{52}$

$P(26\text{...}30 \mid \text{Senior}) = \frac{0}{52} = 0$

$P(46\text{...}50K \mid \text{Senior}) = \frac{40}{52}$

$\therefore P(X \mid \text{Senior}) \propto 0$

$P(X \mid \text{Junior}) \propto 0.0123$

∴ Thus, naive bayes would classify as

Junior ⇒ Ans

3) K-Means:- Distance :- Euclidean

Num_clusters = 2

Initial_clusters: Individual 1 (1.0, 1.0)

Individual 4 (5.0, 7.0)

→ ~~Iter~~ Iteration 1:-

Sr.No.	Points \ Clusters	(1.0, 1.0) 1	(5.0, 7.0) 2
1	(1.0, 1.0)	0.00 ✓	7.211025
2	(1.5, 2.0)	1.11803 ✓	6.10327
3	(3.0, 4.0)	3.6055 ✓	3.6055
4	(5.0, 7.0)	7.211	0.00 ✓
5	(3.5, 5.0)	4.7169	2.5 ✓
6	(4.5, 5.0)	5.315	2.061 ✓
7	(3.5, 4.5)	4.3011	2.9154 ✓

So, Clusters assigned are:-

Cluster 1 → P1, P2, P3

Cluster 2 → P4, P5, P6, P7

→ Iteration 2:-

∴ Updating Cluster heads by taking mean:-

New Cluster Head 1 = (1.833, 2.33)

New Cluster Head 2 = (4.125, 5.375)

P.T.O

Distances are:-

	CH1 (1.833, 2.33)	CH2 (4.325, 5.75)
P1	1.57233 ✓	5.376
P2	0.4714 ✓	4.255
P3	2.0344	1.7165 ✓
P4	5.6396	1.8456 ✓
P5	3.144	0.7288 ✓
P6	3.771	0.53033 ✓
P7	2.73353	1.07529 ✓

Hence, new clusters assigned are

Cluster 1 → P1, P2
Cluster 2 → P3, P4, P5, P6, P7

→ Iteration 3:

Updating Cluster heads:-
New Cluster Head 1 = [1.25, 1.5]
New Cluster Head 2 = [3.9, 5.1]

Distances are:

	CH1	CH2
P1	0.559 ✓	5.021
P2	0.559 ✓	3.9204
P3	3.0516	1.4212 ✓
P4	6.6567	2.19544 ✓
P5	4.1608	0.4232 ✓
P6	4.77624	0.6082 ✓
P7	3.75	0.72111 ✓

Hence, we can see no more change in the clusters, hence K-Means stops.

$$\therefore \quad CH1 = [1.25, \quad 1.5]$$
$$CH2 = [3.9, \quad 5.1]$$

Points in Cluster 1 → P1, P2

Points in Cluster 2 → P3, P4, P5, P6, P7

2>

T.id	Items bought
1	Milk, Tea, Coke
2	Eggs, Tea, Cold Drink
3	Milk, Eggs, Tea, Cold Drink
4	Eggs, Cold Drink
5	Juice

So, Minimum Support count = $0.5 \times 5 = 2.5$

∴ We need Minimum Support count of 3

→ Iteration 1:-

So,

Item	Support
Milk	2 X
Eggs	3
Tea	3
Cold Drink	3
Juice	1 X
Cake	1 X

So, we take

Item	Support
Eggs	3
Tea	3
Cold Drink	3

→ ~~Items bought~~ Iteration 2:-

P.T.O.

Items	Support	
Eggs, Tea	2	×
Eggs, Cold drink	3	
Tea, Colddrink	2	×

∴ We get,

Item	Support
Egg, Cold drink	3

So, Yes, we have 1 itemset

{Eggs, Colddrink} that is a frequent itemset.

4.) Ensembling is a technique of combining two or more algorithms of similar or dissimalar types called base learners. It is done to make a more robust system which incorporates the predictions from all the base learners.

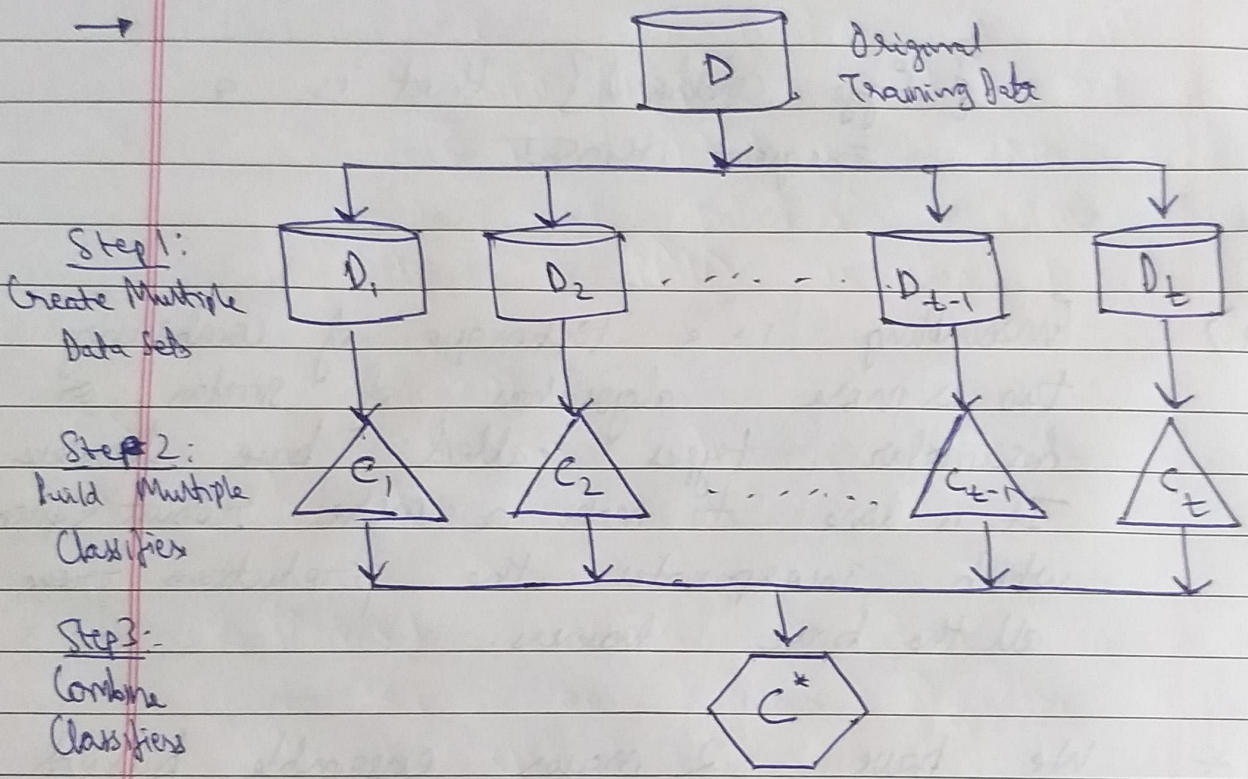
* We have 2 major ensemble based classification methods called as bagging and boosting.

P.T.O.

a) **Bagging:** Bagging is also referred to as bootstrap aggregation. Bootstrapping is a sampling technique in which we choose 'n' observations & rows out of the original dataset of 'n' rows as well.

→ But the key is that each row is selected with replacement from the original dataset so that each row is equally likely to be selected in each iteration.

→



→ We can have multiple bootstrapped samples from the same data. Once we have these multiple bootstrapped samples, we can grow trees for each of these bootstrapped samples and use the majority vote or averaging concepts to get the final

prediction.

Eg: We define 5 bagged decision trees that made the class predictions as yellow, blue, yellow, yellow, red, we would take the most frequent class and predict yellow. This is how a Random Forest works.

b) Boosting:- It was developed to guarantee performance improvements on fitting training data for a weak learner that only needs to generate a hypothesis.

→ Here, instead of sampling, re-weighing examples is done. In boosting, weights are assigned to each training tuple.

→ A series of 'k' classifiers is iteratively learned. After a classifier M_i is learned, the weights are updated to allow the subsequent classifier M_{i+1} to give more attention to the training tuples that were misclassified by M_i.

→ The final boosted classifier, M^* combines the votes of each individual classifier where the weight of each classifier's vote is a function of its accuracy.

→ The boosting algorithm can be extended for the prediction of continuous values. Boosting assigns a weight to each classifier's vote, based on how well the classifier performed.

→ The class with the highest sum is the "winner" and is returned as the class prediction for tuple X.

Eg: We define 5 weak learners for classification of mail as SPAM / HAM and out of these, 3 are voted as 'SPAM' and 2 are voted as 'HAM' and so we take SPAM as classification.

5.) The various clustering methods can be classified as:-

i) K-Means:
Given K, the K-means algorithm applies the following steps:-

- Partition objects into 'k' non-empty subsets
- Compute seed points as the centroids of the clusters of the current partition (mean point)
- Assign each object to the cluster with the nearest seed point.
- Repeat till no new assignment.

i) Partitioning Methods:-

→ Suppose we are given a database of 'n' objects or data tuples, a partitioning method constructs 'k' partitions of the data, where each partition represents a cluster and $k \leq n$.

→ That is, it classifies the data into 'k' groups which satisfy:-
  i) each group must contain at least one object
  ii) each object must belong to exactly one group.

ii) Hierarchical Methods:-

→ Creates a hierarchical decomposition of the given set of data objects.

→ A hierarchical method can be
  a) Agglomerative :- Also called bottom-up approach and starts with each object forming a separate group.
  - It successfully merges the objects or groups that are close to one another, until all of the groups are merged into one or until a termination condition holds.

b) Divisive Approach :- Also called the top-down approach starts with all of the objects in the same cluster.
- Cluster is split up into smaller ones until each point is a single cluster.

ii) Density-Based Methods :-

→ General idea is to continue growing the given cluster as long as the density in the "neighbourhood" exceeds some threshold.

→ The neighbourhood of a given radius has to contain at least a minimum number of points.

eg DBSCAN

v) Grid-Based Methods :-

→ Grid-based methods quantize the object space into a finite number of cells that form a grid structure.
→ All of the clustering operations are performed on the grid structure.
→ Fast processing time.

⇒

vi) Model - based Methods :-

→ Hypothesizes a model for each of the clusters and find the best fit of the data to the given model.

eg EM, COBWEB, SOM.

vii) Constraint - based Method :-

→ Clustering approach that performs clustering by incorporation of user - specified or application - oriented constraints.

→ A constraint expresses a user's expectation & provides an effective means for communicating with the clustering process.