Name of the Subject: DATA ANALYSIS & INFO. EXTR. Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Q3

a

Class.

C1:- buys_Computer = "Yes'

C2:- buys_Computer = 'no'

Data to be classified:-

X = (age <=30, Income = medium, Student = yes, Credit_rating = Fair)

P(C1) = P(buys_Comp = 'Yes') = 9/14
= 0.643

P(buys_Comp = 'no') = 5/14
= 0.357

- Compute P(X | Ci) for each class.

P(age = '<=30' | buys_Comp = 'Yes') = 2/9
= 0.22

P(age = '<=30' | buy_Comp = 'no') = 3/5 = 0.6

P(income = 'medium' | buys_Comp = "Yes') = 4/9 = 0.444

P(income = 'medium' | buys_Comp = 'no') = 2/5 = 0.4

P(Student = 'Yes' | buys_Comp = 'Yes') = 6/9 = 0.667

P(Student = 'Yes' | buys_Comp = 'no') = 1/5 = 0.2

P(Credit_rating = "fair" | buys_Comp = 'Yes') = 6/9 = 0.667

P(Credit_rating = "fair" | buys_Comp = 'no') = 2/5 = 0.4.

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

X = (age <= 30, income = medium, Student = yes, Credit_rating = fair).

P(X | C_i) : P(X | buy_Comp = 'Yes') =
= 0.222 × 0.444 × 0.667 × 0.667
= 0.044

P(X | buys_Comp = 'no') = 0.6 × 0.4 × 0.2 × 0.4 = 0.019

P(X | C_i) * P(C_i) : P(X | buys_Comp = 'Yes') * P(buys_Comp = 'Yes')
= 0.028

P(X | buys_Comp = 'no') * P(buys_Comp = 'no') = 0.007

→ Therefore, X belongs to class ("buys_Comp = Yes")

$\boxed{b}$ Limitation of Naive Bayes

- Naives Bayesian predication requires each requires each Conditional prob. be non-zero.

- Otherwise the predicted prob. will be zero.

$$P(X | C_i) = \prod_{k=1}^{n} P(x_k | C_i)$$

- eg., Suppose a dataset with 1000 tuples, income = low(0) income = medium (990), & income = high(10).

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

- Use Laplacian Correction
- Add 1 to each case.

Prob (income = low) = 1/1003
Prob (income = med) = 991/1003
Prob (income = high) = 11/1003

- The "Corrected" prob estimate are close to their "Uncorrected" ~~con~~ Counterparts.

Q2 b

- K-means, each cluster is represented by the centre of the cluster.

- Given k, the k-means algorithm is implemented in 4 steps.

1. Partition object into k non-empty subsets.

2. Compute seed points as the centroids of the cluster of the current partitioning

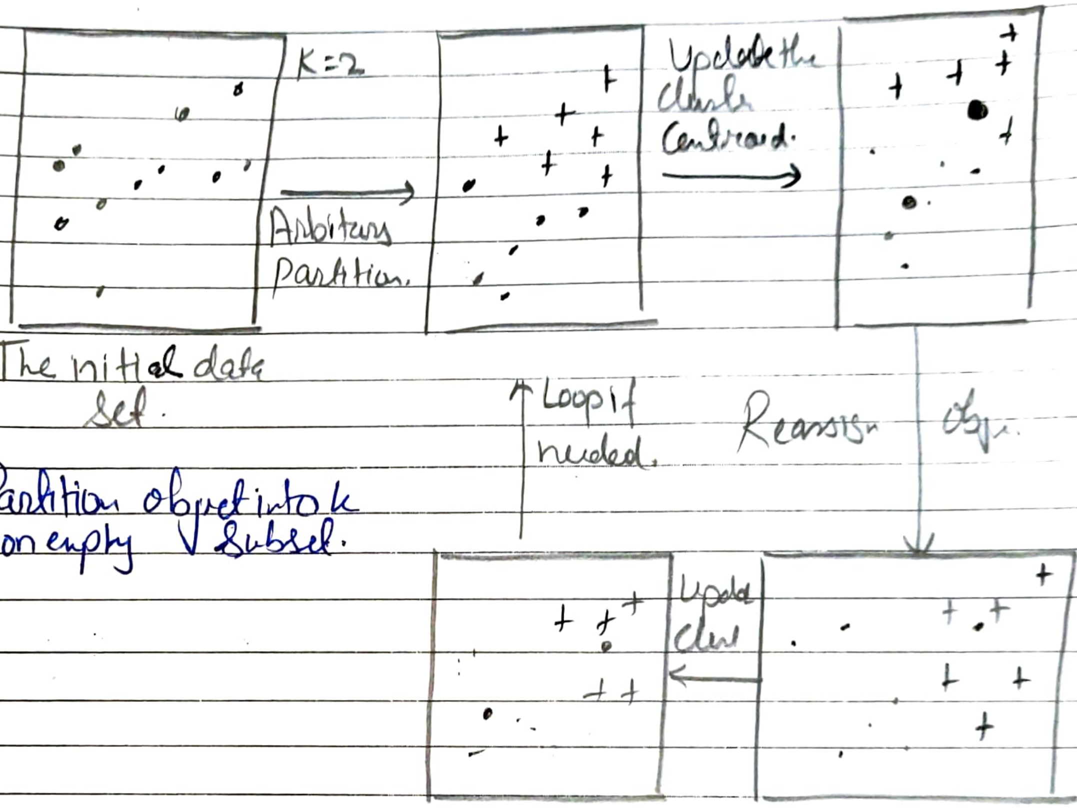
3. Assign each object to the cluster with the nearest seed point.

4. Go back to step 2, stop when the assignment does not change.

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

→ Example.



- Partition object into k non empty subset.

- Limitation.

- K-means clustering Algorithm has limitatio.

1. It requires to specify the no. of cluster (k) in advance

2. It can't handle noisy data or outliers.

3. It is not suitable to identify cluster with non-convex shapes.

Name of the Subject: DAIE Subject Code: IT-704

Seat No: IT076 Student ID: 18ITUBN116 Branch/Sem: IT-VII

Q2 | a | Decision Tree Algorithm. [DTA].

- DTA belongs to the family of Supervised learning algorithm.
- The decision Criteria are different for classification & regression trees.
- Decision Tree use multiple algorithm to decide to split a node into 2 or more sub-node.
- The decision Tree splits selects the ~~splits~~ all nodes available variable & then select the split which result in most homogenous sub-node.
- The algorithm Selection is also based on the type of target variable

ID3, ~~C~~ C4.5, CART (Classification & Regression Tree) CHAID, MARS.