22/10/21

DWM
TT-1

JUNAID GIRKAR
60004190057
TE COMPS A4

Q1
ANS

BINNING :
- considering the neighbourhood of the sorted data smoothening can be applied.
- the sorted data is placed into bins or buckets
- Smoothening by bin means
- Smoothening by bin medians
- Smoothening by bin boundaries.

Different approaches of binning.

a] EQUAL-WIDTH (Distance) PARTITIONING :-
- Divides the range into N intervals of equal size : uniform grid

bin width = (max value − min value) / N

Eg: Consider numbers from 0 to 100.
width of 5 bins = (100 − 0)/5 = 20
∴ Bins formed : [0-20], [20-40], (40-60], (60-80], (80, 100]

b] EQUAL-DEPTH (Frequency) PARTITIONING OR EQUAL-HEIGHT BINNING :
- the entire range is divided into N-intervals, each containing the same number of samples approximately.
- This results in good data scaling
- Handling categorical attributes may be a problem

Example: let us consider sorted data

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

Number of bins (N) = 3

∴ Bin 1 = 4, 8, 9, 15

Bin 2 = 21, 21, 24, 25,

Bin 3 = 26, 28, 29, 34

Q2 STORE 1 [Ascending order]

20, 120, 160, 200, 210, 250, 290, 330, 380, 460, 510, 580

median = $\frac{250+290}{2}$ = 270 = $Q_2$

median of upper half = $\frac{160+200}{2}$ = 180 = $Q_1$

median of lower half = $\frac{380+460}{2}$ = 420 = $Q_3$

IQR (Inter Quartile Range) = $Q_3 - Q_1$ = 240

min = $Q_1 - 1.5$ IQR = 180 − 360 ≤ 0 ∴ min = 0

max = $Q_3 + 1.5$ IQR = 420 + 360 = 780

STORE 2 [Ascending order]

70, 80, 140, 180, 210, 260, 380, 420, 440, 500, 520, 630

median = $\frac{260+380}{2}$ = 320 = $Q_2$.

median of upper half = $\frac{140+180}{2}$ = 160 = $Q_1$

median of lower half = $\frac{440+500}{2}$ = 470 = $Q_3$

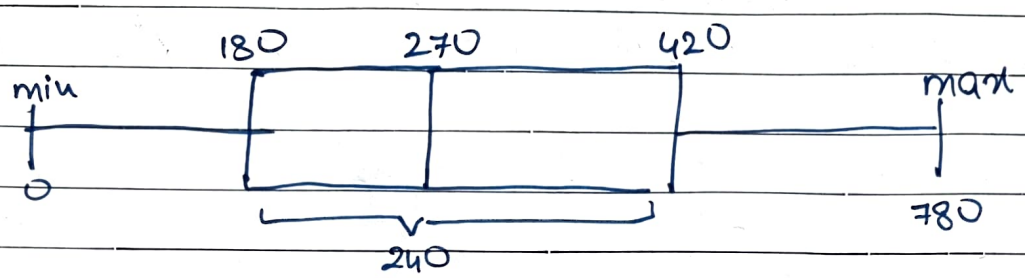
IQR = $Q_3 - Q_1$ = 470 − 160 = 310

min = $Q_1 - 1.5$ IQR = 160 − 465 < 0 ∴ min = 0

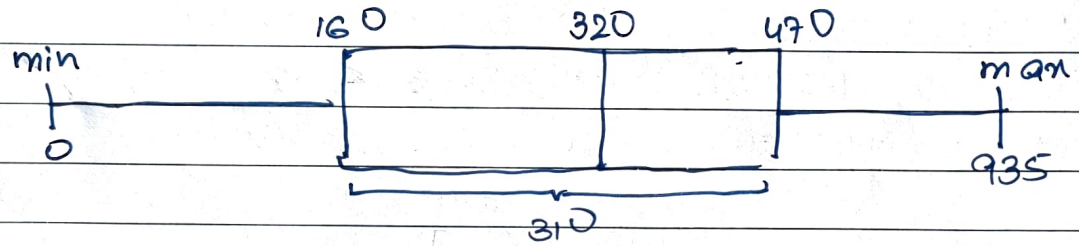
max = $Q_3 + 1.5$ IQR = 470 + 465 = 935

6000419005-7

STORE 1




STORE 2 :




Store 1 has no outliers

Store 2 has no outliers

GOOU190057.

Q3 a Differentiate between bagging and boosting.

| BAGGING | BOOSTING |
|---|---|
| • It is the simplest way of combining predictions that belong to the same type | • It is a way of combining predictions that belong to different types. |
| • Aims to decrease variance, not bias | • Aims to decrease bias, not variance |
| • Each model receives equal weight | • Models are weighted according to their performance |
| • Different training data subsets are randomly drawn with replacement from the entire training dataset. | • Every new subset contains the elements that were misclassified by previous models. |
| • Example : The random forest model uses bagging | • Example : The AdaBOOST uses Boosting techniques |

6000U190057

3b

ANS

- Rule based classification is featured by building rules based on object attributes. Rule based classification is a powerful tool for feature extraction, often performing better than supervised classification for many feature types.
- learned model is represented by a set of IF-THEN rules.
- IF-THEN rule is expressed in the form of IF condition THEN conclusion.
- THE LHS part of the rule is called "rule antecedent" or "pre condition" and RHS part is called as "rule consequent"

- EXAMPLE: IF age = young AND salary = high THEN loan = yes

- This can also be written as:

$$(age = young) \wedge (salary = high) \Rightarrow loan = yes$$

Rule R can be accesed by its coverage and accuracy

$$coverage(R) = n_{covers} / |D|$$

$$accuracy(R) = n_{correct} / n_{covers}$$

where $n_{covers}$ = no of tuples covered by R

$n_{correct}$ = no of tuples correctly classified by R

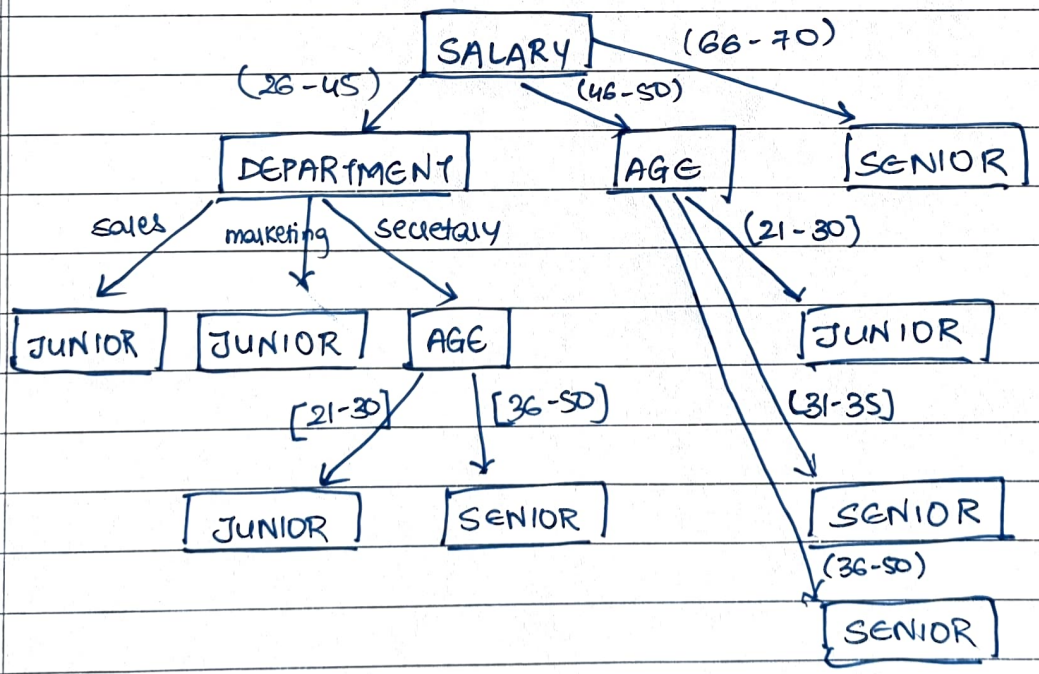
$|D|$ = no of tuples in data set D

- If more than one rule is triggered then it needs conflict resolution and ordering is done based on size.
- Rules are organized based on some measure of rule quality or by taking expert opinion.

6000419005

3C

| DEPARTMENT | STATUS | AGE | SALARY | ~~COUNT~~ |
|---|---|---|---|---|
| sales | senior | 31 - 35 | 46K - 50K | |
| sales | junior | 26 - 30 | 26K - 30K | |
| sales | junior | 31 - 35 | 31K - 35K | |
| systems | junior | 21 - 25 | 46K - 50K | |
| systems | senior | 31 - 35 | 66K - 70K | |
| systems | junior | 26 - 30 | 46K - 50K | |
| systems | senior | 41 - 45 | 66K - 70K | |
| marketing | senior | 36 - 40 | 46K - 50K | |
| marketing | junior | 31 - 35 | 41K - 45K | |
| secretary | senior | 46 - 50 | 36K - 40K | |
| secretary | junior | 26 - 30 | 26K - 30K | |




FINAL TREE.

6000419005 7

Q4 a Discuss the advantages and disadvantages of K-means clustering method

ANS

ADVANTAGES :-

- Relatively simple to implement
- Scales to large data sets
- Easily adapts to new examples
- Generalizes to clusters of different shapes and sizes, such as elliptical clusters

DISADVANTAGES :

- You need to choose 'K' manually using the "loss vs clusters" plot.
- Being dependent on intial values as for a low 'K', you can mitigate this dependence by running K-means several times with different initial values
- K-means has trouble clustering data when clusters are of varying sizes and density.
- Centroids can be dragged by outliers, or outliers might get their own cluster instead of being ignored.

600041900 57

8Q b

| ITEM | A | B | C | D |
|---|---|---|---|---|
| A | 0 | 1 | 4 | 5 |
| B | 1 | 0 | 2 | 6 |
| C | 4 | 2 | 0 | 3 |
| D | 5 | 6 | 3 | 0 |

STEP 1 : let A & B be mediods

Initial clusters are {A, D} & {B, C}


STEP 2 : Determining other 2 points C, D if they replace existing mediods

A replaced by C or D

STEP 2 : Replace non-mediods and compute swapping cost.

Replace A with C

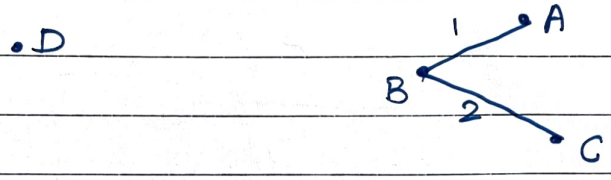
$$Tc_{AC} = C_{ARC} + C_{BAC} + C_{CAC} + C_{DAC}$$

cluster formed :


$$Tc_{AC} = 1 + 0 + (-2) + (-2)$$

$$= -3$$

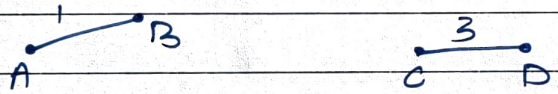
STEP 3: Replace A with D




$TC_{AD} = 1 + 0 + 0 + (-5)$

$= -4$

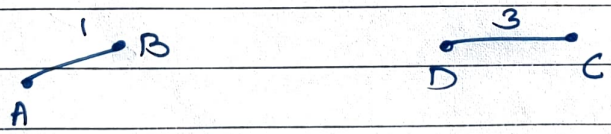
STEP 4: Replace B with C




$TC_{BC} = 0 + 1 + (-2) + (-2)$

$= -3$

STEP 5: Replace B with D




$TC_{BD} = 0 + 1 + 1 + (-5)$

$= -3$

So, the cluster is getting reduced the most in step 3: The choice of final cluster will be: {A, B, C}, {D} as D's values are deterministically far from A, B, C.