Camlin Page
Date / /

2019

(1) a) X (0, -1, 0, 1) Y (1, 0, -1, 0)

$= \sqrt{(-0)^2 + 1^2 + 1^2 + 1^2}$

$= 2$

b) F-measure $= \frac{2rp}{r+p}$

$= \frac{2 \times 0.5 \times 0.6}{1.1}$

$= \frac{0.6}{1.1} = \frac{6}{11}$

Sp 6
10 10
30
100
3
10

c) Yes, {abc} will also be infrequent because according to apriori principle, if an itemset is infrequent, all its supersets are infrequent

d) 1) Eliminate the missing values : we may lose some important attributes req. for analysis

2) Ignore the missing values - may affect our analysis

3) Estimate the missing values - using KNN we can estimate the missing values

e) Precision: The closeness of the measurements with each other is called precision

Bias: The variation of the measurements with the actual value being measured. It is calculated by taking the diff between the mean of all the values and the actual value being measured

f) Variable transformation is the data preprocessing technique which is used to standardize the variables within a given range eg 0-1. If we have a mix of smaller values & larger values we can apply variable transformation and bring them all within a given range. We can deal with outliers as well

Methods:-

1) Normalization or standardization:

Using the Z-score method

$$z = \frac{x - \text{mean}}{\sigma}$$

2) Min-max normalization

$$x' = \frac{x - \min}{\max - \min} (\text{new\_max} - \text{new\_min}) + \text{new\_min}$$

(g) $\text{support}(X \to Y) = \frac{4}{5}$

$\text{conf}(X \to Y) = \frac{3}{4}$

Yes, we can derive the support of $Y \to X$ but not the confidence, as we need the number of occurences of Y.

(h) Advantages :-

i) Whole dataset can be used for testing as well as training.

Disadvantages

i) In Leave one out, one record is used as test record & rest all as training records.

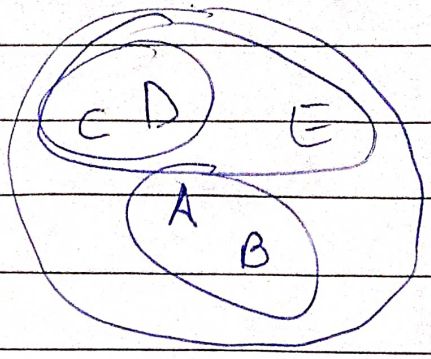
2) So, it is very computational expensive to calculate.

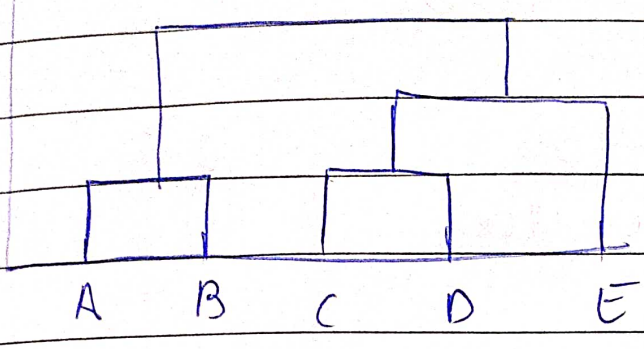
3) It has very high variance

(i)

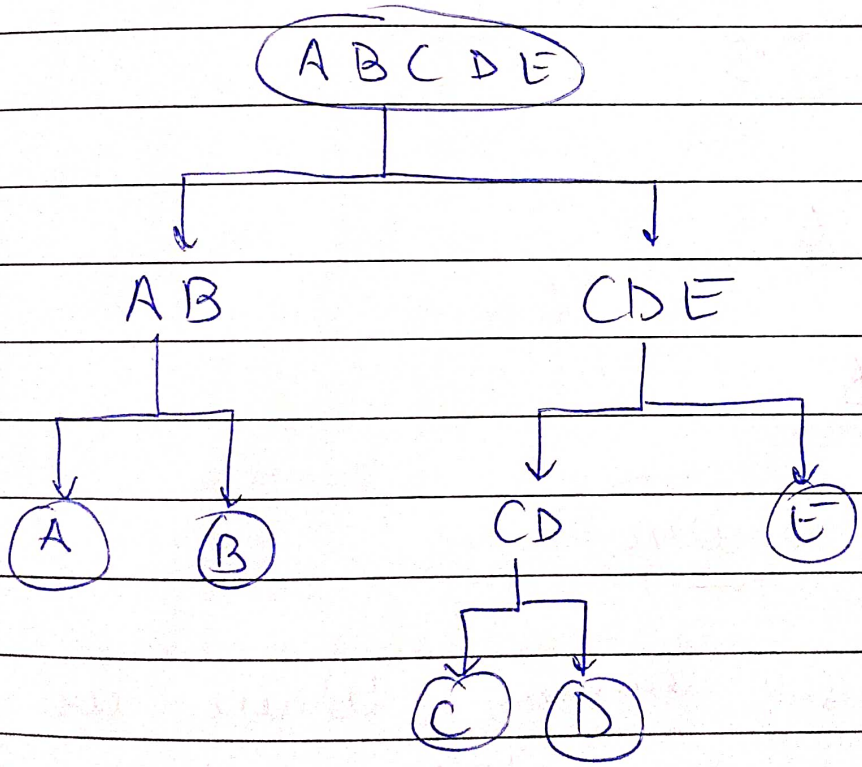
Agglomerative	Divisive
(i) Bottom up approach	(i) Top down approach
(ii) Singleton cluster at first & then successively combine until all clusters are merged into single cluster	(ii) Splitting the cluster until we get the Singleton clusters

Dendogram





Divisive



(J) ** Asymmetric

(K) Accuracy = $\frac{400+300}{1000}$ = 0.7

Sensitivity = $\frac{400}{400+100}$ = 0.8

Specificity = $\frac{300}{300+200}$ = 0.6

TPR = 0.8

FPR = 0.4

SECTION-B

(2) b) ** metric

c) Unsupervised learning :- Where we don't feed in any data and the algorithm on itself classifies the records into diff groups on the basis of their characteristics. Clustering is unsupervised learning eg - Customer segmentation.

Camlin | Page
Date / /

(3) (a) Overall Gini index

(i)

4+ 6-

$$1 - \left[\left(\frac{4}{10}\right)^2 + \left(\frac{6}{10}\right)^2\right]$$

$$= 1 - \left[\frac{4}{25} + \frac{9}{25}\right] \qquad = \boxed{\frac{12}{25}}$$

A10

7T 3F

4+ 3- 3-

T7 $1 - \left[\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right]$

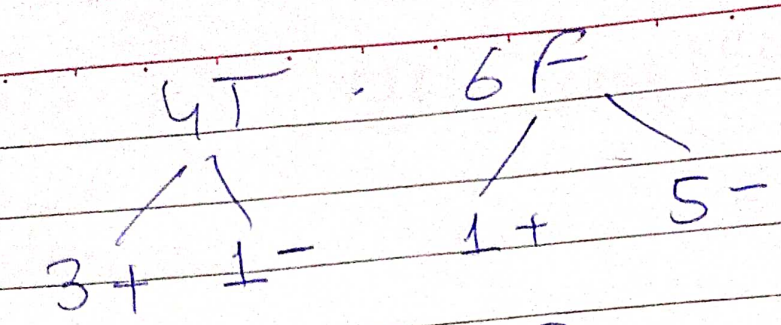
$$= 1 - \left[\frac{16}{49} + \frac{9}{49}\right] = \boxed{\frac{24}{49}}$$

P7 0

Weighted $= \frac{24}{49} \times \frac{7}{10} = \boxed{\frac{12}{35}} = 0.342$

Camlin | Page
Date / /

B



$1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$

$= 1 - \left[\frac{9}{16} + \frac{1}{16}\right] = \frac{6}{16} = \boxed{\frac{3}{8}}$

$1 - \left[\left(\frac{1}{6}\right)^2 + \left(\frac{5}{6}\right)^2\right]$

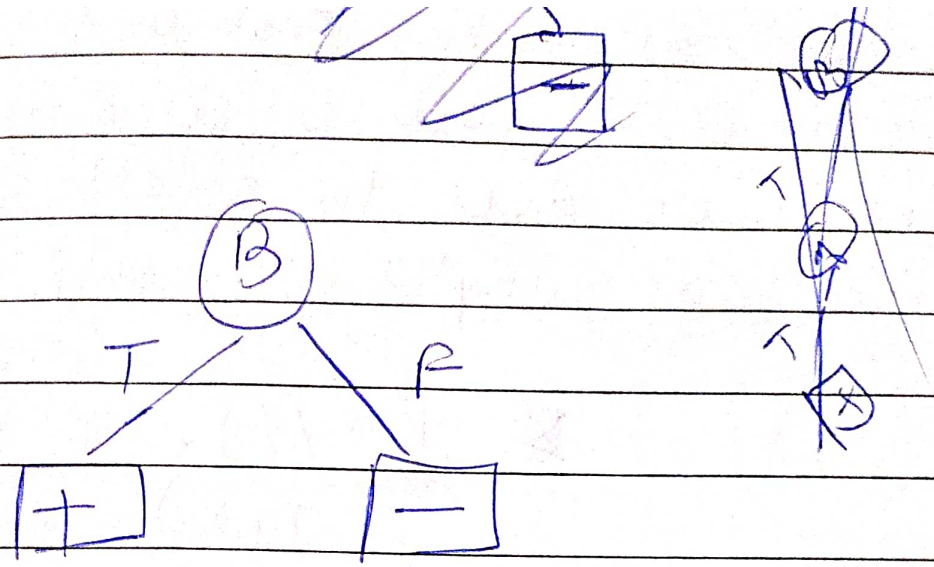
$= 1 - \left[\frac{1}{36} + \frac{25}{36}\right] = \boxed{\frac{5}{18}}$

Weighted $\frac{3}{2\cancel{8}} \times \frac{4}{10} + \frac{\cancel{8}}{3\cancel{18}} \times \frac{\cancel{6}}{\cancel{10}\ 2}$

$= \frac{3}{20} + \frac{1}{6}$

$= \frac{18 + 20}{120} = \frac{38}{120} = \boxed{\frac{19}{60}} = 0.316$

(ii) So, the decision tree induction would choose B.



(iv) 2 instances are misclassified by resulting decision tree

(b) KNN is classified as lazy learner because there is no need to build a model in KNN and it just learns all the training set but any record out of training set comes, it is inefficient to apply

(4) (a) Exhaustive rules: Every record triggers atleast one rule out of the ruleset given. If not exhaustive, we would not be able to classify the record into any class. . . . . Resolved using default class

(b) Progressive sampling: Here, we start with a small sample size & progresses into a large sample size till an optimal size is obtained. Here, we don't need to fix the sample size at start and we can avoid the problems like high variance & low accuracy which can occur due to small sample size and large sample size resp.

(c) $$P(Y|X) = \frac{P(X|Y) \cdot P(Y)}{P(X)}$$

$$\rightarrow P(Y|X) = \frac{\sum_i P(X_i|Y) \cdot P(Y)}{P(X)}$$

Assumption used by Naive Bayes classifier is of conditional independency.

$\Rightarrow$ X is said to be conditionally independent of Y given Z when,

$$P(X|Y,Z) = P(X|Z)$$

11)

5 (a) F-3 itemsets

{1,2,3}
{1,2,4}
{1,2,5}
{1,3,4}
{1,3,5}
{2,3,4}
{2,3,5}
{3,4,5}

F-1 itemsets

{1}
{2}
{3}
{4}
{5}

Using $F_{R-1} \times F_1$

{1,2,3,4}, {1,2,3,5}, {1,2,4,5}, {1,3,4,5}, {2,3,4,5}

(ii) Using $F_{R-2} \times F_{R-2}$

Here, first k-2 elements should be same.

= {1,2,3,4}, {1,2,3,5}, {1,3,4,5}, {1,2,4,5}, {2,3,4,5}

01	0
01	1
10	0

(b) 2 bits

	Binary form
Awful	00
Poor	01
OK	10
Good	11

4 bits

	Binary form
Awful	1000
Poor	0100
Ok	0010
Good	0001

6 (i) $\{e\} = 8 \Rightarrow 80\%$.

$\{b,d\} = 2 \Rightarrow 20\%$.

$\{a,d\} = 4 \Rightarrow 40\%$.

$\{b,d,e\} = 2 \Rightarrow 20\%$.

$\{e\}$ and $\{a,d\}$ are frequent

(ii) $\{b,d,e\}$

Camlin Page
Date Confidence

{b,d}	→ {e}	Conf = 100%
{b,e}	→ {d}	Conf = 2/4 = 50%
{d,e}	→ {b}	2/5 = 40%
{e}	→ {b,d}	2/8 = 25%
{d}	→ {b,e}	2/6 = 33.3%
{b}	→ {d,e}	2/6 = 33.3%

Strong rules →

{b,d} → {e}

b) Nominal - These attributes are used to distinguish the records

eg- eye color, customer ID,

Ordinal : These attributes are used when we want to rank or order the records

eg- grades, positions in race

(7) A 2018 done