# Part – I

(1)

$X = [4.4, 5.1, -3.7, 2.1, -1.9]$

A for $x \le -2.5$

$A = [-3.7]$

B for $-2.5 < x < 2.5$

$B = [-1.9, 2.1]$

C for $x \ge 2.5$

$C = [4.4, 5.1]$

(2) Let assume:

$C_1$ is positive class

$C_2$ is Negative class

| | | Actual class | | |
|---|---|---|---|---|
| | | $C_1$ | $C_2$ | |
| Predicted | $C_1$ | $f_{11}$ | $f_{01}$ | $f_{+1}$ |
| | $C_2$ | $f_{10}$ | $f_{00}$ | $f_{+0}$ |
| | | $f_{1+}$ | $f_{0+}$ | |

example

| | 1 | 0 → Actual |
|---|---|---|
| P 1 | TP | FP |
| 0 | FN | TN |

$$\text{Accuracy} = \frac{\#\text{ Correct}}{\#\text{ total}} = \frac{f_{11} + f_{00}}{f_{11} + f_{01} + f_{10} + f_{00}}$$

Error Rate $\Rightarrow$ $\dfrac{\#\text{ incorrect}}{\#\text{ total}} = \dfrac{f_{01} + f_{10}}{f_{11} + f_{01} + f_{10} + f_{00}}$

FP $\rightarrow$ type I error Reject Null hypothisis.

$FP = \underline{f_{01}} = (f_{+1} - f_{11})$

FN $\Rightarrow$ type II error should have reject Null hyp.

$FN = \underline{f_{10}} = (f_{+0} - f_{00})$

3

Association Rule $\{(A, B) \rightarrow (C)\}$

Confidence $= \dfrac{\cancel{\text{Suppor}}\ S(A, B \cup C)}{S(A, B)}$

$= \dfrac{\sigma((A, B) \cup C)}{\sigma(A, B)}$

Lift (Interest factor) $= \dfrac{S(A, B \cup C)}{S(A, B) \cdot \underline{S(C)}}$

Lift or Interest factor Look into Support of prior et items. while in Confidence we are not considering the support of 'C'. ~~Also~~ Also telling the interesting -ness measure. takes into account by Support of prior ~~prob~~ probability.

Interest factor properties.

$I = 1$ A. B are independent,

$I > 1$ A B positive correlation,

$I < 1$ A.B have Negative correlation.

(4)

we have total n records ∴ K any number

test set $= n/k$

train set (no validation set) $= n - n/k$ ⇒ in this case our $k \to n$

$n/k \to 1$

$\simeq n - n/k \simeq n - 1$

$\simeq n$ (Just less then 'n' like near to 'n').

we can say that our train set will approach to n.

train set $\to n$ #

5 $d = 15$ $N = 12{,}000$

Data matrix size $= n \times d$

$D$

Do

Centered Data matrix $\Rightarrow Z = D - I\mu$

$Z$ dim same $(n \times d)$

Vector of mean of each feature.

Covariance matrix $\Rightarrow$

$$C = \frac{1}{n} Z Z^T \quad \left(\text{Dim} = (n \times d)(n \times d)^T = \underline{n \times n}\right.$$

Normal distribution we can apply PCA.

eigen value decomposition of $C$

$$C = S \Lambda S^T$$

$\Lambda \rightarrow$ eigen value.

$S \rightarrow$ eigenvector matrix

$$\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$$

$Y = PT$

$P$ mystery matrix

$\therefore Y = S^T \cdot D$ or Reduction

$Y = R^T D$

$R \rightarrow$ Subset of $S$

then we can pick 'K' largest eigen value.

where we have good % of total variance.

6

$$X_1 = \overset{x}{(3}, \overset{y}{4)}$$

$$X_2 = \overset{x}{(5}, \overset{y}{12)}$$

manhattan = $|3-5| + |4-12|$

$L_1 = M = 2 + 8 = 10$

euclidean =

$L_2 = E = \sqrt{2^2 + 8^2} = \sqrt{4+64}$

$= \sqrt{68}$

~~Man > e~~

$M > E$

$(L_1 > L_2)$

7

| | $B$ | $\bar{B}$ | |
|---|---|---|---|
| $A$ | $f_{11}$ | $f_{10}$ | $f_{1+}$ |
| $\bar{A}$ | $f_{01}$ | $f_{00}$ | $f_{0+}$ |
| | $f_{+1}$ | $f_{+0}$ | $N$ |

Lift (Interest factor)

$$= \frac{P(A,B)}{P(A)\,P(B)}$$

$$= \frac{N\, f_{11}}{f_{1+} \cdot f_{+1}}$$

(8)

$$\phi = \frac{f_{11} f_{00} - f_{01} f_{10}}{\sqrt{f_{1+} f_{+1} f_{0+} f_{+0}}}$$

eg =

| | $B$ | $\bar{B}$ | |
|---|---|---|---|
| $A$ | 60 $f_{11}$ | 10 $f_{10}$ | 70 |
| $\bar{A}$ | 10 $f_{01}$ | 20 $f_{00}$ | 30 |
| | 7 | 30 | 100 |

if we invariant with out changing the value then :

$f_{00} \to f_{11}$ & $f_{01} \to f_{10}$

basically $0 \to 1$

Still over $\phi$ coff same ~~tat~~ in this case.
#

Part II

1) $$\cos\theta = \frac{\vec{A} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|}$$ → find the correlation between two vectors.

$$= \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}$$

$\cancel{A} = (3, 4, 5) \qquad B = (5, 12, 13)$

$$\cos\theta = \frac{\cancel{\Sigma}\ \cancel{15}(3 \cdot 5 + 4 \cdot 12 + 5 \cdot 13)}{\sqrt{\cancel{(9^2 + 16^2}\ 3^2 + 4^2}}$$

$$\cos\theta = \frac{(3 \times 5 + 4 \times 12 + 5 \times 13)}{\sqrt{(3^2 + 4^2 + 5^2) \times (5^2 + 12^2 + 13^2)}}$$

Part - 2

2

| | | Actual 1 | Actual 0 |
|---|---|---|---|
| P | 1 | TP | FP |
| | 0 | FN | TN |

(Sensitivity)

$$\text{Recall } (r) = \frac{TP}{TP + FN}$$

if we want high Recall (r) then our model should have low number of FN.

$$\uparrow r = \frac{TP}{TP + FN \downarrow}$$

High Recall means more value of reality (I mean more predict positive).

$$\text{Recall (TPR) True positive Rate} = \frac{TP}{TP+FN}$$

(3) Part 2

$$\text{min sup} = \frac{60}{100} \times 6 = 3.6$$

| a, b, c | setep1 | sup |
|---|---|---|
| a c | a | 3 ⟶ x |
| b c | b | 3 ⟶ x |
| a | | |
| b | c | 4 ⟶ |
| c | | |

In Step 1 we got C as frequent item only.

$$\text{Support } (a \to c) = \frac{\sigma(a \cup c)}{N} = 2/6 = \underline{\underline{1/3}}$$

$$\text{Confidence } c = \frac{\sigma(a \cup c)}{\sigma(a)} = \underline{\underline{2/3}} \quad \left(R = 3^3 - 2^4 + 1 = 12\right)$$

Since at min Sub = 3.6 (60%) in this case we have C as frequent only. So we can not make any valid Rule for only.

# Part II

(4)

e.v = [ 35, 25, 20, 15, 5 ]

percent of variane explained by each e.v.

= total = 35 + 25 + 20 + 15 + 5

= 100

% of variane explaine = [35%, 25%, 20%, 15%, 5%]

we want ~~...~~ reduce dimensionality by 80%. that means we wave to add 3 e.v.

= 35% + 25% + 20% = 80%

to 3 dim from (5 = d) has been reduce.

S.d = $\sqrt{var}$ = Sd $PC_1$ = $\sqrt{35}$ #

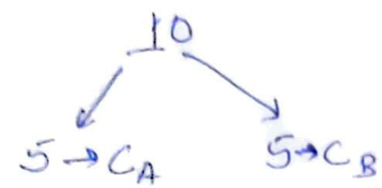
Sd $PC_2$ = $\sqrt{25}$ #

Sd $PC_3$ = $\sqrt{20}$ #

Part 3

1.

Decision tree containing 10 records.




$P(C_A) = 5/10 = 0.5$

$P(C_B) = 5/10 = 0.5$

misclassification = $1 - \sum_{i=1} max$

$= 1 - max(P(i|t))$

$= 1 - 0.5 = 0.5$ (befor split)

Gini = $1 - \sum_{i=0}^{i=1} P(i|t)^2$

$= 1 - ((0.5)^2 + (0.5)^2)$

$= 1 - \frac{2}{4}$

$= 0.5$ (befor split).

Entropy $\rightarrow$ $$-\sum_{i=0}^{i-1} p(i|t)\log_2 p(i|t)$$

$$= -\frac{1}{2}\underbrace{\log_2(1/2)}_{-1} - \frac{1}{2}\underbrace{\log(1/2)}_{-1}$$

$$= 1 \text{ (befor split).}$$

- Measure purity/impurity befor and after the split.
  - K # of children in split.

$$\Delta = \underset{\text{node befor split}}{I(\text{parent})} - \sum_{i=1}^{K} \frac{N(v_j)\; I(v_j)}{N}$$

$I(\,) \rightarrow$ impurity function

$N \Rightarrow$ ~~num~~ number of total parent records

$N(v_j) \rightarrow$ number of records at child.

After optimal split $\rightarrow$

that means we do not have any ~~p~~ missclassification error.

Missclassification Rate $= 1 - \max(5/5, 0/5)$

$= 1 - 1 = 0$ #

Gini at child ~~not~~ node

$$G_A = 1 - \sum_{i=1}^{b-1} (P(i|t))$$

$$= 1 - (1 + 0)$$

$$= 0 \#$$

Entropy at optimal childnode.

$$-\sum_{i=0}^{i-1} P(i|t) \log_2 P(i|t)$$

$$-1 \log_2 1 - 0 \log_2 0$$

$$\downarrow \qquad \downarrow$$

$$0 \qquad 0$$

$$= 0$$

$\#$

**Lucky 7** – Bonus Questions (Industry News, AI/ML Topics) – 1 point each, 7 points total

1. What model recently released by DeepMind allows for accurate prediction of 3-dimensional shape of a protein molecule given input amino acids?

   AlphaFold

2. Which firm recently fired its head of AI ethics, shortly after the controversial departure of one of its senior researchers?

   Google fired Margaret Mitchell

3. What family of algorithms were recently developed which are able to solve classic treasure hunting video games such as Pitfall on Atari?

   Go-explore

4. What disease was IBM able to predict the onset of based on changes in writing/language via the use of machine learning models?

   Alzheimer's disease

5. What category of modified videos did a consortium led by Facebook/Microsoft/Cornell/MIT recently introduce a detection challenge for?

   Deepfakes

6. Which firm recently released a new image recognition algorithm that was trained on over 1 billion images, but did not require manual labels?

   Facebook

7. What quantum computing goal was recently achieved by Google which was revealed to the public via NASA?

   Quantum supremacy