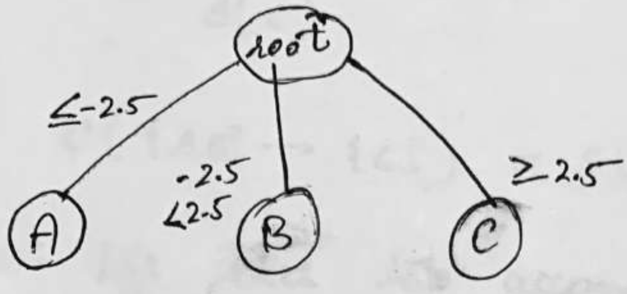
Data Mining

Short Answers (Part I)

1.



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

Ans:

$A = \{-3.7\}$

$B = \{2.1, -1.9\}$

$C = \{4.4, 5.1\}$

~~2. classes $\{C_1, C_2\}$~~

	C_1	$\overline{C_1}$	
C_2	f^{11}	f^{10}	f^{1+}
$\overline{C_2}$	f^{01}	f^{00}	f^{0+}
	f^{+1}	f^{+0}	N

2. Classes $\{C_1, C_2\}$

		Actual Class		
		C_1	C_2	
Predicted class	C_1	$TP = f^{00}$	$FP = f^{10}$	f^{+0}
	C_2	$FN = f^{01}$	$TN = f^{11}$	f^{+1}
		f^{0+}	f^{1+}	

$TP = f^{00}$

$FP = f^{10}$

$FN = f^{01}$

$TN = f^{11}$

$TP + FP = f^{+0}$

$FN + TN = f^{+1}$

$TP + FN = f^{0+}$

$FP + TN = f^{1+}$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{f^{00} + f^{11}}{(f^{+0}) + (f^{+1})}$$

$$\text{Error rate} = \frac{FP + FN}{TP + TN + FP + FN} = \frac{f^{01} + f^{10}}{f^{+0} + f^{+1}}$$

3. For $\{A, B\} \longrightarrow \{C\}$

$$\text{Lift} = \frac{c(\{A,B\} \longrightarrow \{C\})}{s(C)} = \frac{s(\{A,B,C\})}{s(A,B)\ s(C)}$$

$$c(\{A,B\} \longrightarrow \{C\}) = \frac{s(A,B,C)}{s(A,B)}$$

Ans: Lift takes into account the support for Consequent that lift doesn't. This gives us a better value which tells us the ~~corett~~ correlation betten the Association Rule.

4. data set size $= n$

test set $= \frac{n}{K}$

training set $= n - \frac{n}{K}$

total trainset size approached,

$$= 2^{n} - 2$$

5. $d = 15$

$N = 12{,}000$

dimension of covariance matrix $= 15 \times 15$

for multivariate normal (Gaussian) distribution, number of distribution parameters wolld be $= 15 \times 15$

6. $x = [3,4]$ $\quad x_2 = [5,12]$

$$L_1 = |x_2 - x_1| + |y_2 - y_1|$$
$$= |5-3| + |12-4| = 2+8 = 10$$
$$L_1 = 10$$

$$L_2 = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2} = \sqrt{2^2 + 8^2} = \sqrt{4+64}$$
$$L_2 = \sqrt{68} = 8.246211$$

$$(L_1, L_2) = (10,\ 8.246211)$$

Ans: values will be larger under L_1 norm.

7. Contingency table:

	B	$\bar{B}$	
A ~~B~~	f^{11}	f^{10}	f^{1+}
$\bar{A}$ ~~$\bar{B}$~~	f^{01}	f^{00}	f^{0+}
	f^{+1}	f^{+0}	N

$$\text{Lift} = \frac{P(A,B)}{P(A)\,P(B)} = \frac{f^{11}/N}{(f^{1+}/N)\cdot(f^{+1}/N)}$$

$$\text{Lift} = \frac{N f^{11}}{f^{1+} \times f^{+1}}$$

8. By definition, ϕ-coefficient for binary variables is given as

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

Ans: As in numerator we use f^{00} (i.e null addition factor), upon adding unrelated data f^{00} will increase and thus ϕ will not be invariant

(Part II) Long Answers

1.

$x = [3, 4, 5]$

$y = [5, 12, 13]$

$$\text{Cosine similarity} = \frac{x \cdot y}{|x|\ |y|}$$

$$= \frac{(5\times3) + (12\times4) + (13\times5)}{\sqrt{3^2+4^2+5^2}\ \sqrt{5^2+12^2+13^2}}$$

$$= \frac{15 + 48 + 65}{\sqrt{9+16+25}\ \sqrt{25+144+169}}$$

$$= \frac{128}{5\sqrt{2} \cdot 13\sqrt{2}} = \frac{128}{130}$$

$$= 0.98461538$$

2.

Recall is defined as,

$$\text{Recall} = \frac{TP}{TP + FN}$$

- Highest value of recall can be 1 and is only possible when FN = 0, as any value of FN is adding weight to denominator which will reduce the recall value.

- A simple model which achieves the maximum value for ~~sta~~ recall can be ~~a~~ implemented by minimizing the value of FN in the model.

3. Transaction

$\{a, b, c\}$

$\{a, c\}$

$\{b, c\}$

$\{a\}$

$\{b\}$

$\{c\}$

minsup = 60%

$s(a) = 3/6 = 50\%$

$s(b) = 3/6 = 50\%$

$s(c) = 4/6 = 66\%$

$s(a,b) = 1/6 = 16\%$

$s(a,c) = 2/6 = 33\%$

$s(b,c) = 2/6 = 33\%$

$s(a,b,C) = 1/6 = 16\%$

Ans: only itemset $\{c\}$ will be frequent.

$$s(\{a\} \longrightarrow \{c\}) = s(a,c) = 33\%$$

$$c(\{a\} \rightarrow \{c\}) = \frac{s(a,c)}{s(a)} = \frac{2/6}{3/6} = \frac{2}{3} = 66.66\%$$

Given the minimum support (min sup) value of 60%, rule $\{a\} \longrightarrow \{c\}$ will not be valid as support for $\{a\}$ is ~~term~~ less than the minsup value, thus their superset $\{a, c\}$ will be pruned and thus its association rule $\{a\} \rightarrow \{c\}$ is not possible.

4. Given eigen-value are [35, 25, 20, 15, 5]

Thus cumulative values of PCA are,

PCA 1 = 35%

PCA with 2 components = 35 + 25 = 60%

~~PC~~ PCA with 3 components = 35 + 25 + 20 = 80%

∴ The Analyst ~~not~~ would be able to reduce from 5 dimensions to 3 dimensions for selection of 3 PCAs.

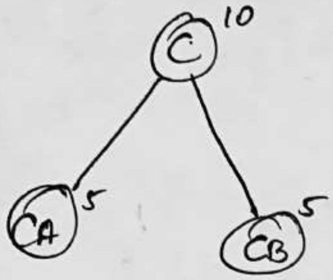
$$\sigma(PCA1) = \sqrt{\frac{35}{100} \times 100} = \sqrt{35} = 5.916$$

$$\sigma(PCA2) = \sqrt{\frac{25}{100} \times 100} = \sqrt{25} = 5$$

$$\sigma(PCA3) = \sqrt{\frac{20}{100} \times 100} = \sqrt{20} = 4.4721$$

Part III (Essay Question)

1.



$P(CA) = 5/10$
$P(CB) = 5/10$

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

$= 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2 = 1 - 0.25 - 0.25$

$= \underline{0.5}$

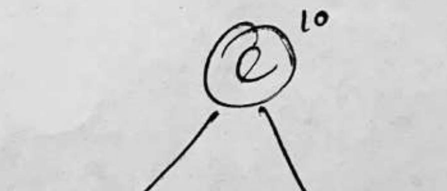
entropy $= -\left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right) - \left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right)$

~~= 0.69316~~ $= 1$

Misclassification error $= 1 - \max[\cancel{0}5/10, 5/10]$

$= 1 - 0.5 = \underline{0.5}$

Optimal Case :



$P(CA) = 0$
$P(CB) = 1$

gini $= 1 - (0)^2 - (1)^2 = \underline{0}$

Entropy $= -0\log_2(0) - 1\log_2(1) =$

$= -0 - 1.0 = \underline{0}$

Misclassification error $= 1 - \max[0/10, 10/10]$

$= 1 - 1 = \underline{0}$

Thus for equal split we get maximum possible impurity in each impurity measure. But for optimal split we get minimum value i.e 0 for all.

Lucky 7 (Bonus Questions)

1. Alphafold

2. Tmnit Gebru, Emily Bender were the involved entities. Firm is Google.

3. Go - explore

4. Alzheimer's

5. Deepfake video detection is the challenge.

6. Facebook is the firm that recently released the new image recognition algorithm trained over 1 billion images.

7. Quantum Supremacy was achieved in tasks that would usually take days, this milestone was recently achieved by Google and revealed to people via. NASA.