20231_dsci_552_32416 Midterm

Saketh Reddy Regatte

TOTAL POINTS

62.5 / 110

QUESTION 1

Decision Trees and Bias/Variance Dilemma 42 pts

1.1 **1 / 6**

- **0 pts** Correct
- **1 pts** X-axis label missing or incorrect or too generic. Example label: Number of tree nodes.
- **1 pts** Y-axis label missing or incorrect
- **1 pts** Missing or wrong labels for bias and variance curves
- **3 pts** Missing or wrong bias and variance curves
- **3 pts** Wrong explanation of bias/variance dielemma
- **1 pts** Missing total error curve
- **1 pts** Explanation should be specific to decision trees

✓ **- 5 pts** *Should draw three error curves bias, variance and total as a function of # of tree nodes*

1.2 **0 / 6**

- **0 pts** Correct
- **1 pts** Relationship: Test error is like the total error in the bias/variance diagram
- **1 pts** Relationship: Train error is like the bias error in the bias/variance diagram
- **1 pts** Test error should be approximately U-shaped
- **1 pts** Train error should start high then approximately continuously decrease
- **6 pts** Not plots of train and test error curves

✓ **- 6 pts** *Blank*

1.3 **0 / 4**

- **0 pts** Correct
- **4 pts** Blank
- **3 pts** Reversed overfitting and underfitting region
- **2 pts** Missing underfitting region

✓ **- 4 pts** *Incorrect labeling*

1.4 **6 / 6**

✓ **- 0 pts** *Correct*

- **3 pts** don't list the two types of tree pruning
- **1 pts** don't mention overfitting in purpose
- **2 pts** No description of two types of tree pruning
- **3 pts** list the wrong types of tree pruning
- **2 pts** answer the wrong purpose of tree pruning
- **3 pts** don't mention the purpose
- **1 pts** don't mention the prepruning and postpruning
- **1 pts** no description of two types of tree pruning

1.5 **2 / 8**

✓ **- 6 pts** *wrong calculation and conclusion*

- 8 pts wrong calculation and conclusion

- 2 pts wrong calculation of internal node's cost

- 0 pts Click here to replace this description.

- 8 pts no calculation and wrong conclusion

- 1 pts There is no discussion about the size of N in the final conclusion (hint: if Tree 1 cost == Tree 2 cost -> N = ?)

- 2 pts no conclusion (substitute 16 -> d and 4 -> k into the equation)

- 2 pts wrong conclusion

- 4 pts wrong calculation and conclusion

1.6 **8 / 8**

- 2 pts Few missing instances

- 4 pts Almost Half Instances missing / Half correct

- 6 pts Very few instances shown

- 8 pts Blank / Little to no instances shown

✓ **- 0 pts** *Correct*

1.7 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts No / Wrong answer

QUESTION 2

Density Estimation 16 pts

2.1 **4 / 4**

✓ **- 0 pts** *Correct*

- 2 pts Missing $$p_g$$ in equation. The correct equation should be $$\prod_{t=1}^{N} (1-p_g)^{x^t-1}*p_g$$

- 1 pts Missing $$x^t-1$$ in equation. The correct equation should be $$\prod_{t=1}^{N} (1-p_g)^{x^t-1}*p_g$$

- 4 pts wrong equation, The correct equation should be $$\prod_{t=1}^{N} (1-p_g)^{x^t-1}*p_g$$

- 3 pts Missing product from 1 to N. The correct equation should be $$\prod_{t=1}^{N} (1-p_g)^{x^t-1}*p_g$$

- 1 pts Should not include log. The correct equation should be $$\prod_{t=1}^{N} (1-p_g)^{x^t-1}*p_g$$

- 8 pts Wrong form of tree

2.2 **2 / 8**

- 0 pts Correct

✓ **- 4 pts** *Wrong derivative process.*

- 1 pts Missing $$x^t$$

✓ **- 2 pts** *Wrong final equation $$p_g = \frac{N}{\sum_{t=1}^{N} x^{t}}$$*

- 1 pts Missing N in final equation $$p_g = \frac{N}{\sum_{t=1}^{N} x^{t}}$$

- 2 pts Wrong derivative answer

- 1 pts Wrong form of $$l(p_g)$$

- 1 pts wrong final answer

2.3 **4 / 4**

✓ **- 0 pts** *Correct*

- 4 pts Wrong equation, the prior should be $$ p(p_g|x) = \frac{p(X|p_g)p(p_g)}{p(X)} $$

- 1 pts Missing final equation, the prior should be $$ p(p_g|x) = \frac{p(X|p_g)p(p_g)}{p(X)} $$

QUESTION 3

Clustering 14 pts

3.1 **2 / 8**

- **0 pts** Correct
- **6** *Point adjustment*

Partially correct approach. 23 is not the right answer. Negative coordinates not considered.

3.2 **0 / 6**

- **0 pts** Correct
- **6** *Point adjustment*

Incorrect

- **1 pts** Incorrect Similarity / No answer given
- **1.5 pts** Incorrect Distance Metric Isomap / No answer given
- **1.5 pts** Incorrect Distance metric LE / No answer given
- **0.5 pts** Partially correct distance metric Isomap

✓ - **0.5 pts** *Partially correct distance metric LE*

- **0.5 pts** Partially correct similarity
- **4 pts** No correct answer given

It uses Gaussian Kernel constructed from neighbours within a given distance which may be calculated using local distance metrics like mahalanobis

QUESTION 4

Dimension Reduction 12 pts

4.1 **4 / 4**

- **4 pts** No answer / Incorrect answer

✓ - **0 pts** *Represents the variance captured by the ith component*

- **2 pts** Partly correct

4.2 **2 / 4**

- **0 pts** Correct
- **2 pts** Wrong / Missing Similarity

✓ - **2 pts** *Wrong / Missing Difference*

- **1 pts** Partially right similarity
- **1 pts** Partially right difference
- **4 pts** Completely incorrect answer / Blank

4.3 **3.5 / 4**

- **0 pts** Correct Distance Metric for Isomap and Laplacian Eigenmaps. Correct Similarity.

QUESTION 5

Naive Bayes Classification 12 pts

5.1 **7 / 8**

✓ - **0 pts** *Correct*

- **2 pts** Multiplication with prior probabilities is missing
- **4 pts** Calculation of P("each word" | Spam) is not correct
- **1 pts** Final answer is missing
- **1 pts** Calculation of prior probabilities is missing
- **7 pts** Incorrect
- **2 pts** Formulae are missing, just randomly multiplied the numbers
- **4 pts** P(No) not calculated, final answer missing and P("each word| No) is not calculated properly
- **4 pts** P(No) not calculated, final answer missing and P("each word| No) is not calculated.
- **2 pts** P(X| C0) not calculated properly

- **6 pts** Incorrect
- **1 pts** P(Card | both instances (spam and not spam) not calculated properly
- **1 pts** Calculation error (12/125 x 5/8)=0.06
- **1 pts** P(Deal|~Spam) is incorrect and the final calculation also
- **1 pts** Final answer incorrect
- **1 pts** P(Card | not spam) not calculated properly
- **4 pts** Formulae incorrect
- **4 pts** Solved only half of the problem
- **2 pts** Calculation of P("each word" | No) is not correct
- **5 pts** Incorrect

- 1 *Point adjustment*

It should be P(Credit Card Deal | No) > P(Credit Card Deal | Yes)

5.2 **4 / 4**

- **3 pts** don't mention promotion is not in the dataset

✓ **- 0 pts** *Click here to replace this description.*

- **0 pts** Click here to replace this description.
- **2 pts** don't mention promotion is not in the dataset
- **4 pts** no description
- **4 pts** promotion is not in the dataset
- **2 pts** use credit and card to classify
- **2 pts** wrong conclusion

QUESTION 6

Association Rules 14 pts

6.1 **4 / 4**

✓ **- 0 pts** *Correct*

- **4 pts** Incorrect
- **2 pts** Calculation Error

6.2 **4 / 4**

✓ **- 0 pts** *Correct*

- **4 pts** Blank
- **2 pts** Calculation error
- **4 pts** Incorrect

6.3 **1 / 6**

- **0 pts** Correct

✓ **- 5 pts** *Incorrect*

- **2 pts** P(Interest) > P(Credit, Interest) is not mentioned
- **4 pts** Mentioned only formlae correctly.
- **2 pts** Did not mention the formulae
- **1 pts** Did not mention the final answer
- **6 pts** Incorrect
- **1 pts** Did not mention the other formula

Name: SAKETH REDDY REGATTE
Reg No: 410136705

DSCI 552 MIDTERM

2 March 2023

For this exam one page of notes is allowed (both sides).

Calculators are allowed, but not smartphones, laptops or any device with internet connection.

The exam is 2 hours long and it is for 110 points. **You get a bonus of 10 points!**

There are 6 problems and 20 pages total.

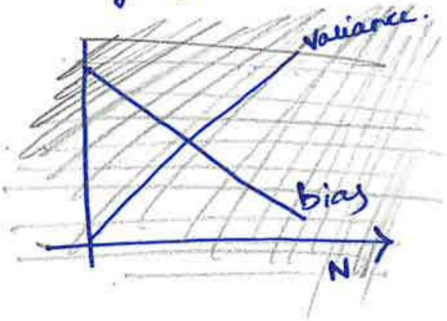
Please remember to write your name

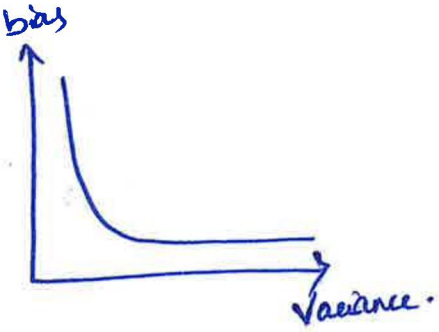
Problem	Points
1	_____/42
2	_____/16
3	_____/14
4	_____/12
5	_____/12
6	_____/14
Total	_____/110

1. (42 points) Decision Trees and Bias/Variance Dilemma
   a. (6 points) Explain the bias/variance dilemma **specifically** in the context of **decision trees**. Draw a diagram of bias/variance to illustrate your explanation. Be sure to carefully label each part of your diagram.

Bias/Variance trade off w.r.t decision trees.

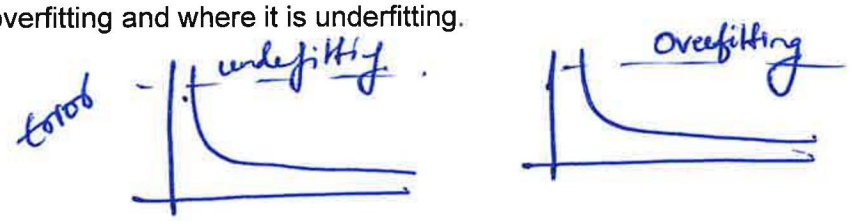
- When number of training instances increases, the bias decreases whereas the variance increases.
- In terms of decision trees training instances are the number of features.





   b. (6 points) Draw a diagram of train **and test error** curves that should be typical of decision trees. What is the relationship between train and test error curves to the curves in the bias/variance diagram?

c. (4 points) For the diagram in part b label the region where the decision tree is overfitting and where it is underfitting.



d. (6 points) What is the purpose of tree pruning? Describe the two types of tree pruning.

• The Purpose of Tree pruning is to avoid the problem of overfitting of the decision tree.

Two types of Pruning Techniques:-

① Prepruning :- In this method the tree {decision nodes} are pruned {stopped generating} before the construction of the tree. It follow the hill climbing method with complexity $O(d^2)$.

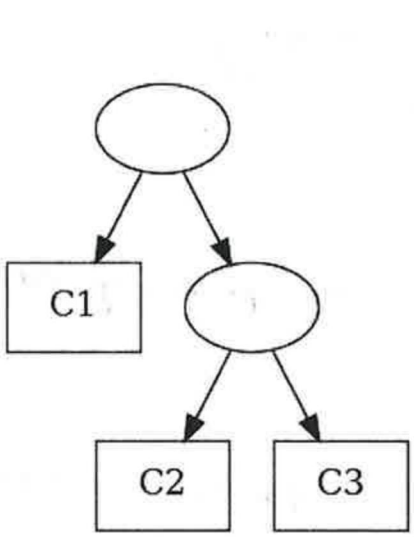
② Post pruning :- This is a backtracking approach where the decision tree is constructed and then overfitted nodes are pruned.

* The purpose of the pruning is to stop the decision tree algorithm to make model simpler (Occam's Razor) which in turn avoid the problem of overfitting.

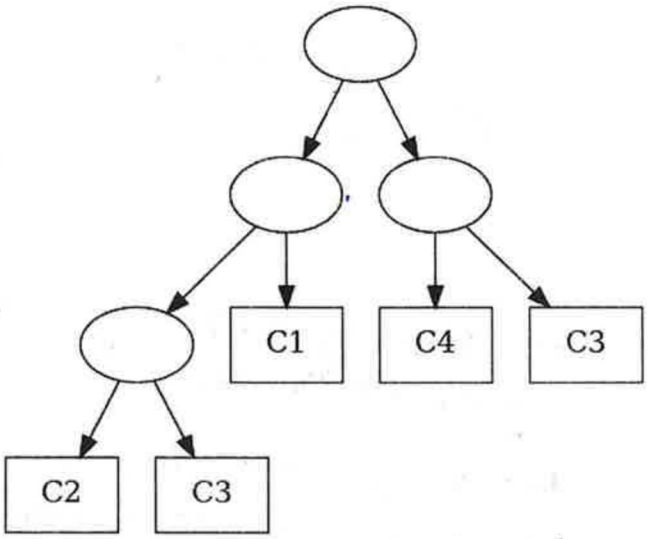
e. (8 points) Minimum description length (MDL) principle.
Consider the two decision trees below. Assume they are generated from a dataset of 16 binary attributes and 4 classes, C_1, C_2, C_3 and C_4. Assume

- Each internal node is coded using $log_2 d$ bits, where d is the number attributes.
- Each leaf node is encoded using $log_2 K$ bits where K is the number of classes.
- For simplicity assume the cost of encode a tree is the total cost of encoding the internal nodes and leaf nodes.
- Each error is encoded using $log_2 N$ bits, where N is the number of training instances.

According to MDL principle which decision tree is better as a function of N?



Tree 1 with 8 errors



Tree 2 with 4 error

According to the Minimum Description Language (MDL).

$\text{Cost}(\text{Model}/\text{Data}) = \text{Cost}(\text{Data}/\text{Model}) + \text{Cost}(\text{Model})$.

Tree 1

$MDL = \log_2 d_1 + \log_2 k_1 + \log_2 N$.

$f_1(N) = \log_2 2 + \log_2 4 + 8 \log_2 N$

$= 3 + 8 \log_2 N$

$f_2(N) = \log_2 4 + \log_2 4 + 4 \log_2 N$

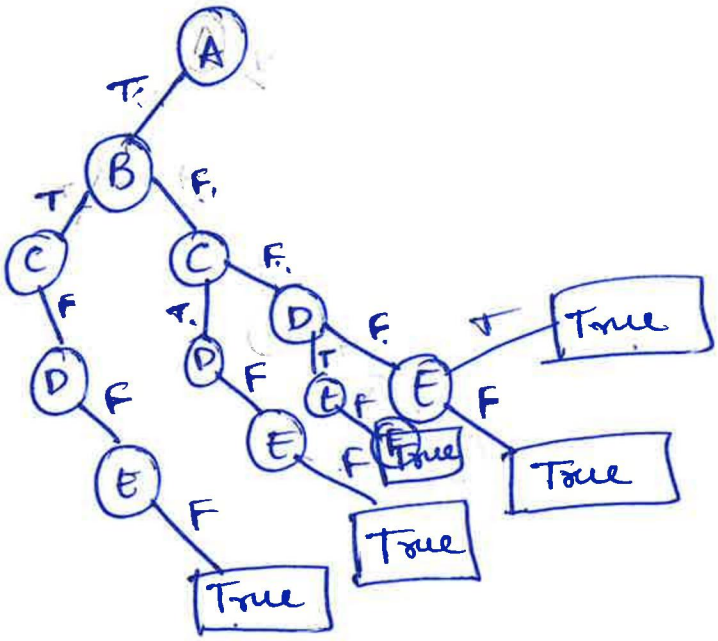
$= 4 + 4 \log_2 N$

$f_1(N) > f_2(N)$

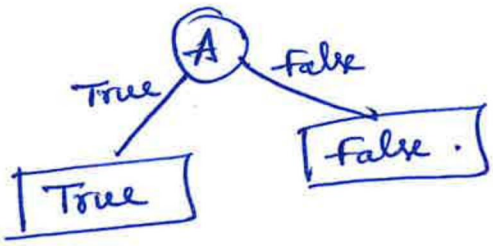
∴ Tree 2 is better compared to tree 1.

f. (8 points) Domingos (2012) points out that overfitting can be caused by noise, but bad learning algorithms can also cause overfitting. For the Boolean training dataset below, draw a decision tree that will **only** classify correctly the positive instances in the training dataset and **no other positive instances** (it will ignore all negative instances).

A	B	C	D	E	Class
T	F	F	F	F	T
T	T	F	F	F	T
T	F	T	F	F	T
T	F	F	T	F	T
T	F	F	F	T	T
F	T	T	T	T	F
F	F	F	F	F	F
F	T	F	F	F	F
F	F	T	F	F	F
F	F	F	T	F	F
F	F	F	F	T	F



g. (4 points) Using the dataset in the previous part, draw the smallest decision tree that will classify the entire dataset correctly with zero training error, i.e. without considering the **no other positive instances** restriction.



2. (16 points) Density estimation

a. (4 points) An entomologist is studying the behaviors of dung beetles by collecting a dataset of the number of attempts individual dung beetles need to successfully push a ball of dung uphill. The dataset collected is a dataset of N beetles $X = \{x^t\}$, where beetle t failed on the first $x^t - 1$ attempts, and succeeded on the last attempt. The entomologist assumes the beetles are not intelligent enough to learn across attempts, so he uses a geometric distribution $p(x) = (1 - p_g)^{x-1} p_g$, where p_g is the probability of success. Write down the **likelihood** equation for parameter p_g.

$$\mathcal{L}(p_g/x) = \prod_{t=1}^{N} p_g * (1-p_g)^{x^t - 1}$$

b. (8 points) Derive **maximum likelihood estimate** of p_g.

log likelihood $\Rightarrow$ $\log(\mathcal{L}(p_g/x))$.

$$= \log\left(\prod_t p_g * (1-p_g)^{x^t-1}\right).$$

$$= \sum_t \left[\log p_g \cdot (1-p_g)^{x^t-1}\right].$$

$$= \sum_t \left[\log p_g + (x^t - 1)\log(1-p_g)\right].$$

$$= N \log p_g + \log(1-p_g)\left[\sum_{t=1}^{N}(x^t - 1)\right].$$

$$\frac{\partial L}{\partial p_g} = 0$$

$$\Rightarrow \quad 0 = N/p_g + \frac{1}{(1-p_g)}\left[\sum x^t - N\right]$$

$$\Rightarrow \quad N/p_g = \left[N - \sum x^t\right] * \frac{1}{1-p_g}.$$

$$\Rightarrow \quad N(1-p_g) = p_g\left[N - \sum x^t\right].$$

$$\Rightarrow \quad (1-p_g) = p_g\left[1 - \frac{\sum x^t}{N}\right]$$

$$\Rightarrow \quad 1 - p_g = p_g - m \cdot p_g \Rightarrow$$

$$\Rightarrow \quad p_g(1+m) = 1$$

$$p_g = \frac{1}{2-m}$$

c. (4 points) To the surprise of the entomologist the beetles in this dataset only needed about half the number of attempts as reported in entomology literature. Suppose the entomologist was able to obtain the prior density from literature. Write down the equation the entomologist needs to solve to incorporate the **prior density**.

Given prior density is known

Acc to bayes theorum

$$P(C_i/x) = \frac{P(x/C_i) \cdot P(C_i)}{P(x)}$$

Prior

3. (14 points) Clustering
   a. (8 points) Show K-mean clustering is not robust to outliers. Consider this one-dimensional dataset of 6 instances $X = \{0, 1, 2, 10, 11, 12\}$. For K=2 clusters add one outlier to the dataset that will cause the K-mean clustering to place the outlier in its own cluster, and the rest of the dataset in the other cluster. What is the closest location this outlier can be to the other points in the dataset, and still be in its own cluster?

• Consider seed points as {0,1}.

Considering Manhattan distance (or) CityBlock distance = $|x_1-x_2| + |y_2-y_1|$

	0	1	
0	0	1	C_1
1	1	0	C_2
2	2	1	C_2
10	10	9	C_2
11	11	10	C_2
12	12	11	C_2
23	23	22	C_2

∴ Kmean clustering is not robust to outliers.

As we calculate the distance and centroid for each iteration, the outlier is clustered as a true value. To avoid this the datapoint should be atleast (23)

   b. (6 points) Outlier detection. Consider these two functions:
      - $d_k(x)$: the distance to the k-th nearest neighbor to instance x
      - $ave_k(x)$: the average $d_k(n)$ over n, where instance n is in the set of the k nearest neighbor of instance x

      Describe how to combine these two functions to use it for outlier detection, where k is a hyperparameter that we can change. Use the dataset in part a. to describe your solution.

* $ave_k(x) = \frac{Sum(x)}{\#(x)}$.

$d_k(x)$: distance to the k^{th} nearest neighbour to instance x.

if $ave_k(x) > d_k(x)$ "Outlier"

else $ave_k(x) \leq d_k(x)$ "Not outlier"

4. (12 points) Dimension Reduction

a. (4 points) In Principal Component Analysis (PCA) what does the eigenvalue λ_i of the ith component represent?

- The eigen value of the i^{th} component expresses the variance along the i^{th} component.

b. (4 points) What are the similarities and differences between Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)?

Similarities

- Both PCA and LDA projects the instances to lower dimensions
- Both are linear dimensionality reduction techniques.
- Both utilize the covariance matrix to project the instances.

Differences:

LDA considers the attributes for the dimensionality reduction to project to lower dims but ~~LDA~~ PCA doesnot considee attributes

c. (4 points) Describe the distance metrics used by Isomap and Laplacian Eigenmaps. What is similar about these two metrics?

- Distance metrics used by ISOMAP: Geodesic distance

  It will considee the local maxima and calculate the distance for each iteration.

- Laplacian Eigenmaps Distance =

  $d^2 = (x_1 - x_2)^T C^{-1} (x_1 - x_2)$.

  ↳ inverse of covariance matrix.

  Called as Mahalobian distance.

∴ both the metrics will work on the distance measure.

5. (12 points) Naive Bayes Classification
   a. (8 points) Use the Naive Bayes assumption and the dataset table below to classify the words: **Credit Card Deal**. Show your work, not just the final answer.

Words	SPAM
Interest Free Card	No
Cash Credit Gift	Yes
Mortgage Interest Deal	No
Cash Back Credit Card	No
Debt Free Deal	No
Credit Card Interest	No
Exclusive Free Deal	Yes
Card Interest Mortgage	Yes

3/8

5/8

* Lets follow the approach of bag of words. Count the frequency of each word

Naive Bayes classifier states that

$$C_{NB} = \underset{c_i \in C}{\arg\max}\ P(c_i) \cdot P(x_1/c_i)\, P(x_2/c_i) \cdots P(x_n/c_i).$$

where $P(c_i) =$ # of instances classified as c_i / size (X).

$$P(x_i/c_i) = \frac{\text{\# of instances classified as } c_i \text{ and has value } x_i}{\text{\# of instances classified as } c_i}$$

$$P(\text{Credit, Card, Deal}/\text{Yes}) = P(\text{Yes}) \cdot P(\text{Credit}/\text{Yes}) \cdot P(\text{Card}/\text{Yes}) \cdot P(\text{Deal}/\text{Yes})$$
$$= 3/8 \cdot 1/3 \cdot 1/3 \cdot 1/3 = 0.014$$

$$P(\text{Credit, Card, Deal}/\text{No}) = P(\text{No}) \cdot P(\text{Credit}/\text{No}) \cdot P(\text{Card}/\text{No}) \cdot P(\text{Deal}/\text{No})$$
$$= 5/8 * \frac{2}{5} * \frac{3}{5} * \frac{2}{5}$$
$$= 0.06$$

$$P(\text{Yes}) = \frac{0.014}{0.014 + 0.06} = 0.19$$

$$P(\text{No}) = \frac{0.06}{0.014 + 0.06} = 0.81$$

$P(\text{No}) > P(\text{Yes})$

∴ Credit Card Deal is Not Spam.

b. (4 points) Describe how you would classify the words: **Credit Card Promotion**.

$P(\text{Credit, Card, Promotion} / \text{Yes})$

$= P(\text{Yes}) \cdot P(\text{Credit}/\text{yes}) \cdot P(\text{Card}/\text{yes}) \cdot P(\text{Promotion}/\text{yes})$

$= 3/8 \cdot 1/3 \cdot 1/3 \cdot 0/3.$

Now, Promotion is not part of the dataset

So, Consider Laplace Estimator μ.

where μ can be any constant $\mu = 1$.

Now, $3/8 \cdot \frac{1+\mu}{3} * \frac{1+\mu}{3} * \frac{0+\mu}{3}$

$= 3/8 \cdot 2/3 * 2/3 \cdot 1/3.$

$= 0.056.$

$P(\text{credit, card, promotion} / \text{No})$

$= 5/8 * 2/5 \cdot 3/5 * \boxed{0/5}$

again adding laplacian estimator with $\mu = 1$ we get

$= 5/8 \cdot \frac{2+1}{5} * \frac{3+1}{5} * \frac{0+1}{5}$

$= 5/8 \cdot 3/5 \cdot 4/5 \cdot 1/5.$

$= 0.3$

$P(\text{Yes}) = \frac{0.056}{0.056+0.3}$ $\quad P(\text{No}) = \frac{0.3}{0.056+0.3}$

$= 0.16$ $\quad = 0.84$

∴ Again Credit Card Promotion is categorized as not spam.

6. (14 points) Association rules

Use the dataset in Question 5. Given the association rule: **Interest → Card**

a. (4 points) What is the support of this rule?

By definition: $\text{Support} = \frac{\text{freq}(\text{Interest}, \text{Card})}{N} = \frac{3}{8} = 0.375.$

$\therefore$ Support = 37.5%

b. (4 points) What is the confidence of this rule?

By definition: $\text{Confidence} = \frac{\text{freq}(\text{Interest}, \text{Card})}{\text{freq}(\text{Interest})} = \frac{3}{4} = 0.75.$

$\therefore$ Confidence (Interest → Card) = 75%.

c. (6 points) Show why if this rule has low confidence:

Credit Interest → Card —①

Then this rule can be pruned:

Interest → Card Credit —②

freq 1 item set

Credit	3
Interest	4
Card	4.

freq (credit, Interest) = 1

freq (credit Interest, Card) = 1

freq (Interest) = 4

freq (Interest, Card, Credit) = 1

$\therefore$ Confidence (Credit Interest → Card) $= \frac{\text{freq}(\text{credit Interest, Card})}{\text{freq}(\text{credit, Interest})}$

$= \frac{1}{1} = 100\%.$

$\therefore$ Confidence (Interest → Card Credit) $= \frac{\text{freq}(\text{Interest, Card credit})}{\text{freq}(\text{Interest})}$

$= 1/4 = 25\%.$

If confidence of ① is low ⇒ freq (Credit, Interest) is high. which reduces the confidence of (Interest → Card Credit) as the denominator freq (Interest) will increase. Because of low confidence ② can be pruned.

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>