Name: Jonathan Farrell Kwuma
Reg No: 498795042

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!**

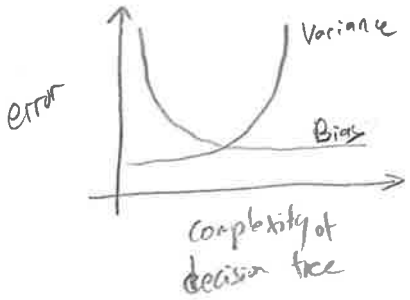
There are 6 problems and 20 pages total.

Please remember to write your name

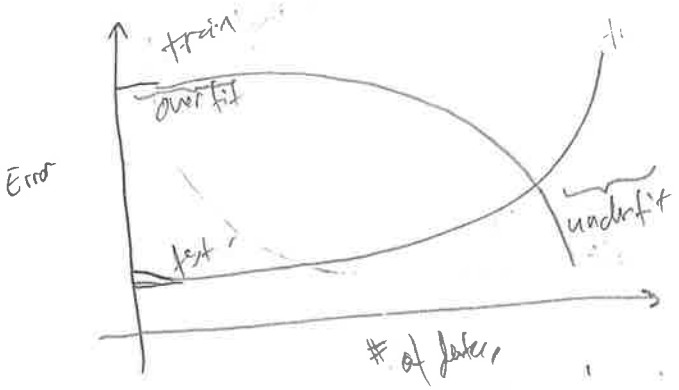
Problem	Points
1	_____/42
2	_____/16
3	_____/14
4	_____/12
5	_____/12
6	_____/14
Total	_____/110

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.

The more complex the decision tree model is, the lower the bias, however the variance would be higher. Meanwhile, the less complex the model is, the lower the variance but higher bias.



   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?



the higher the train test, the higher the over. more training data increases Variance and decreases bias.

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

shown in (B)

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

Prepruning: Stopping the tree earlier

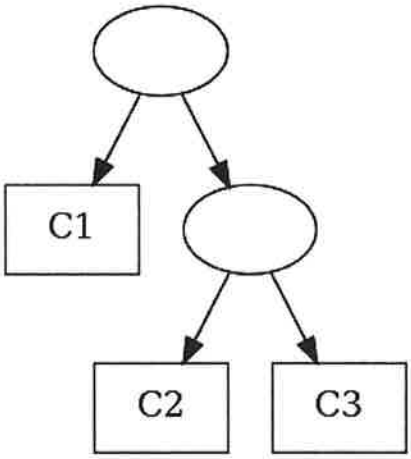
Post pruning: Grow the whole tree then prune the subtrees that overfit on the pruning set

Purpose: Prevent overfitting for better generalization (decrease variance)

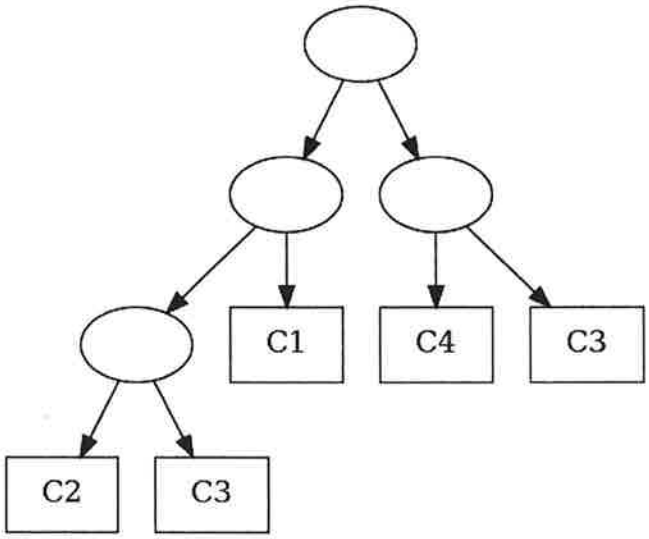
e. (8 points) Minimum description length (MDL) principle.
Consider the two decision trees below. Assume they are generated from a dataset of 16 binary attributes and 4 classes, C_1, C_2, C_3 and C_4. Assume

- 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

Tree 1 because lower computing cost.

MDL: find simple model that provides the most concise description of the data, quantifies model complexity in terms of the number of bits required to encode the data.

$\log_2(6) + \log_2(2) \log_2(3) = 5.585$

$\log_2(16) = 4$

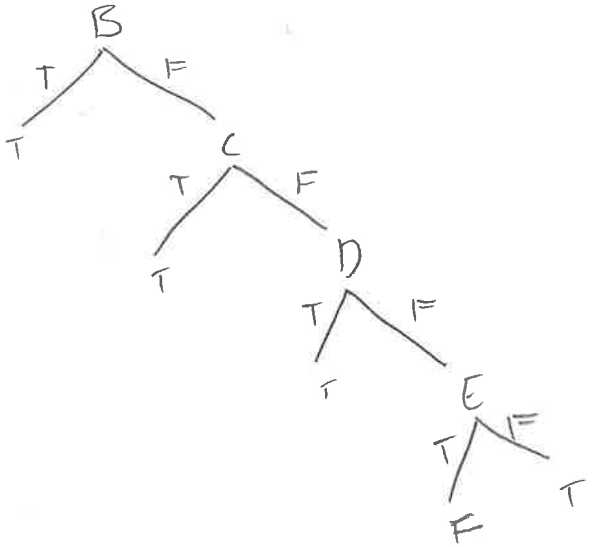
$\log_2(4) = 2$

$\log_2(4) \log_2(4) \log_2(5) = 6.322.$

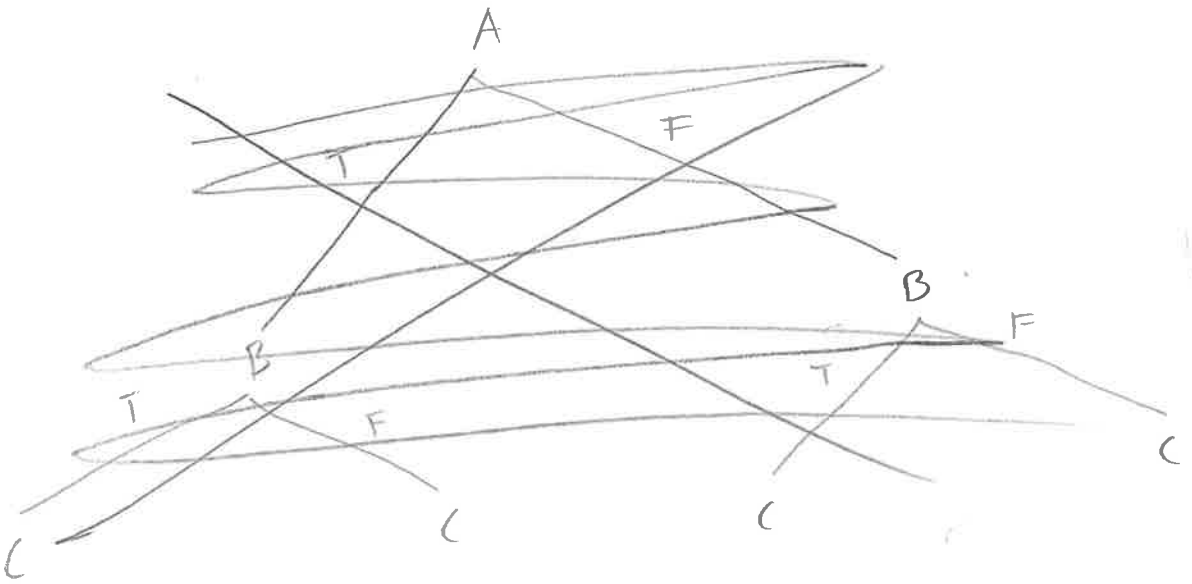
$5.585 < 6.322$

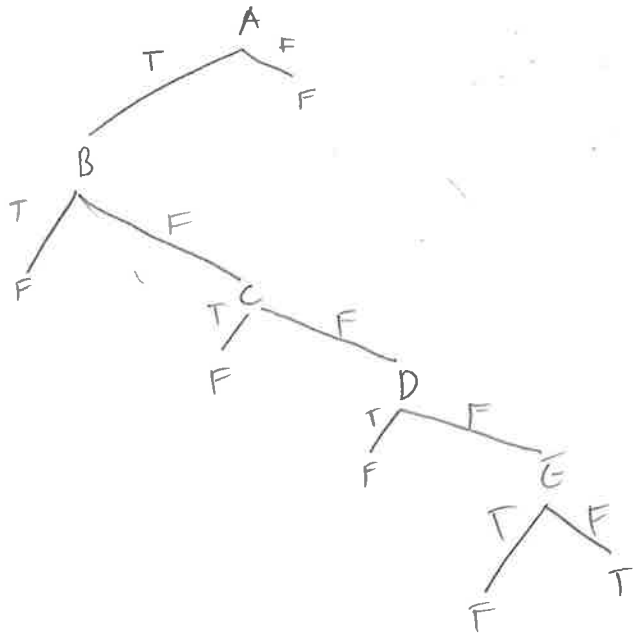
f. (8 points) Domingos (2012) points out that overfitting can be caused by noise, but bad learning algorithms can also cause overfitting. For the Boolean training dataset below, draw a decision tree that will **only** classify correctly the positive instances in the training dataset and **no other positive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

$$l = \prod_{t=1}^{N} p_g \cdot (1-p_g)^{(x-1)}$$

$$L = \log \prod (1-p_g)^{x-1} p_g$$

$$= \sum (x-1) \log(1-p_g) + \log(p_g)$$

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

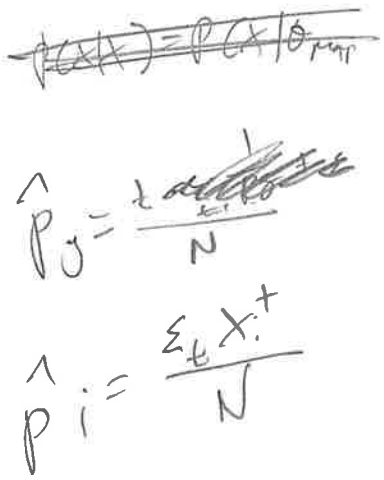
$$\frac{dP}{dx} = \cancel{\log(1-p_g)}$$

$$\frac{dL}{dx} = \sum 1 - \sum \frac{1}{p_g} + \sum \frac{1}{p_g} = 0$$

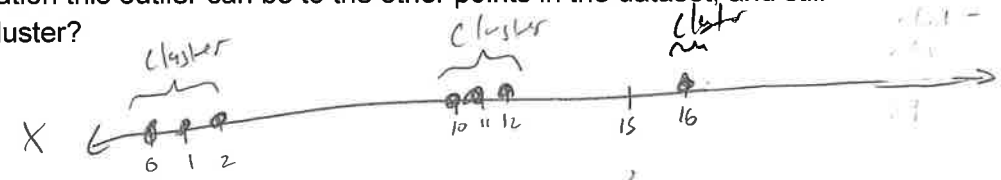
log likelihood function.

finding this derivative equals to zero. finds the maximum of the likelihood estimate.

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**.



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?



if 15 will cluster together

must be >#15

must be >15

$\frac{1+1+x}{3} = 2$

$1+1+x = 6$

$x = 4$

$1+4 = 5$

15 is the closest

4 from the centroid

   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.

Use the $d_k(x)$ to calculate all the mean distance and use the $ave_k(x)$ to compare if the newly calculated mean distance with the new point is higher or lower. If the mean distance is higher, then the point is likely an 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 eigenvalue λ_i is the distance from the points it is the projection directions in a covariance matrix of PCA.

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

PCA is an unsupervised dimension reduction that learns the projection directions that maximize data variance while LDA is a supervised learning method used for classification & dimensionality reduction.

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

distance metric Isomap = Geodesic distance
distance metric Laplacian Eigenmaps = uses a dot matrix of neighborhood preserving embeddings.

Both metrics are low-dimensional representation of data.

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

CCD

S → $\frac{3}{8}$

NS → $\frac{5}{8}$

$P(\text{Credit} \mid \text{Spam}) = \frac{1}{3}$

$P(\text{card} \mid \text{Spam}) = \frac{1}{3}$

$P(\text{deal} \mid \text{Spam}) = \frac{1}{3}$

$P(\text{Credit} \mid \text{NS}) = \frac{2}{5}$

$P(\text{card} \mid \text{NS}) = \frac{3}{5}$

$P(\text{deal} \mid \text{NS}) = \frac{2}{5}$

$P(S \mid CCD) =$

$P(NS \mid CCD) =$

$\frac{P(CCD \mid S)\,P(S)}{P(CCD)} =$

$\frac{P(CCD \mid NS)\,P(NS)}{P(CCD)} =$

$\frac{1}{3} \times \frac{3}{8} + \frac{1}{3} \times \frac{3}{8} + \frac{1}{3} \times \frac{3}{8} = 0.375$

$\frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} = 0.037$

$\frac{2}{5} \times \frac{3}{5} \times \frac{2}{5} = 0.096$

$0.096 > 0.037$

~~$\frac{2}{5} \times \frac{3}{8} + \frac{3}{5} \times \frac{2}{5} \times \frac{2}{5} = 0.875$~~

~~$0.875 > 0.375$~~

So Credit Card Deal will be classified as Not Spam.

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

find the & compare $P(\text{credit}|\text{spam}) \times P(\text{spam}) + P(\text{card}|\text{spam}) \times P(\text{spam}) + P(\text{Promotion}|\text{spam}) \times P(\text{spam})$ and for spam

$\Rightarrow P(\text{credit}|\text{N spam}) \times P(\text{N spam}) + P(\text{card}|\text{N spam}) \times P(\text{N spam}) + P(\text{promotion}|\text{N spam}) \times P(\text{N spam})$ for Not spam

the higher probability will be classified with the class associated with the equation this uses the Bayes theorem.

* We can't use Naive Bayes because $P(\text{promotion}|\text{spam})$ is zero.

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?

Support (Interest → Card)

= P(Interest, Card) = # of "interest" and "card" / # data.

= 3/8 = 0.375

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

Confidence (Interest → Card)

= P(Card | Interest)

= 3/4 = 0.75.

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

Credit Interest → Card

Then this rule can be pruned:

Interest → Card Credit

Confidence (Credit Interest → Card)

= P(Card | Credit Interest) = 0/4

P(Interest | Credit Card) = 2

because confidence of "Credit interest" is zero

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>