20231_dsci_552_32416 Midterm

Harsh Parikh

TOTAL POINTS

72 / 110

QUESTION 1

Decision Trees and Bias/Variance Dilemma 42 pts

1.1 **5 / 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

Bias and variance errors are functions of tree size, not error threshold

1.2 **2 / 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 **4 / 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 **5 / 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 **4 / 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 **4 / 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 **3 / 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 **6 / 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 **0 / 8**

- **- 0 pts** Correct

- 8 *Point adjustment*

Incorrect. C1 and C2 shall become {0,1,2,10,11,12} and {20} respectively. 20 considered as an outlier without an explanation. Limits for outlier not mentioned

3.2 **0 / 6**

- **- 0 pts** Correct

- 6 *Point adjustment*

No Attempt

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 **4 / 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 **4 / 4**

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

- **- 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

QUESTION 5

Naive Bayes Classification 12 pts

5.1 **4 / 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

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 **3 / 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

- 1 *Point adjustment*

Final answer should be it should be pruned because of lower confidence

Name: Harsh Jinesh Parikh
Reg No: 2348 6131 52

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.

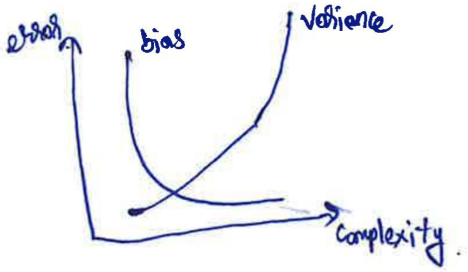
Please remember to write your name

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

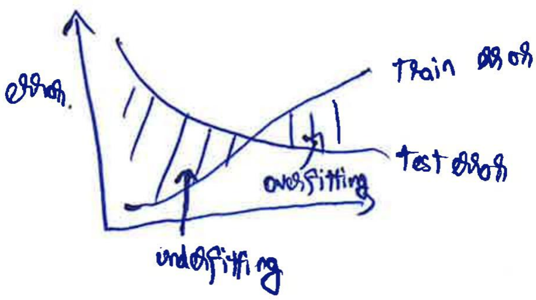
1. (42 points) Decision Trees and Bias/Variance Dilemma
   a. (6 points) Explain the bias/variance dilemma **specifically** in the context of **decision trees**. Draw a diagram of bias/variance to illustrate your explanation. Be sure to carefully label each part of your diagram.

As complexity of decision tree increases, variance increases & bias decreases

Higher variance leads to overfitting.



   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.

Underfitting - test error > train error
Over fitting - train error > test error
(From diagram b))

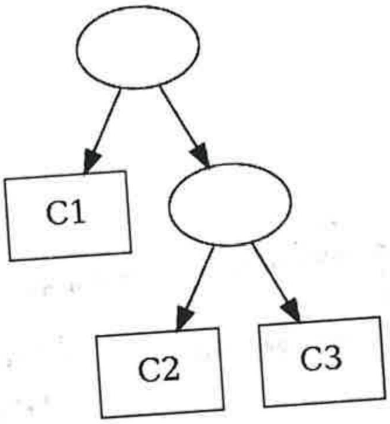
d. (6 points) What is the purpose of tree pruning? Describe the two types of tree pruning.

The purpose of tree pruning is to reduce the ~~generalization~~ error caused due to the subtrees. As ~~For increasing~~ generalization increases, variance decreases. (It removes generalized subtrees).
The two types of tree pruning are -
* Prepruning - which is early stopping of the tree creation hence faster.
* Postpruning - Grow the entire tree then prune subtrees that overfit, hence more accurate.

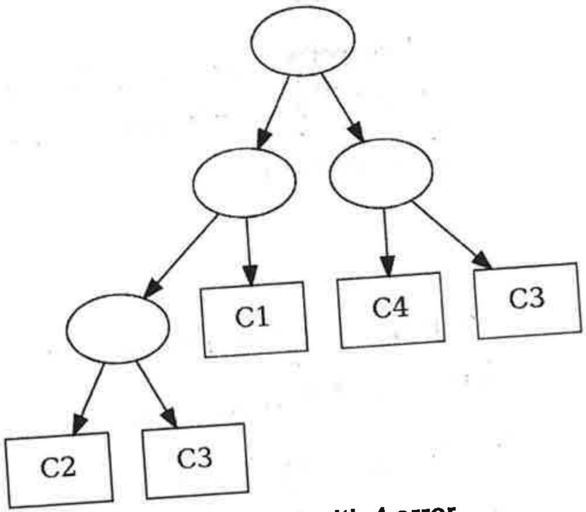
e. (8 points) Minimum description length (MDL) principle.
Consider the two decision trees below. Assume they are generated from a dataset of 16 binary attributes and 4 classes, C_1, C_2, C_3 and C_4. Assume

- Each internal node is coded using $log_2 d$ bits, where d is the number of 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

MDL:- $Cost(M,D) = Cost(D|M) + Cost(M)$

where, cost → no. of bits for encoding., $M \rightarrow$ Model, $D \rightarrow$ Data

$$Cost(Tree1, data) = 8(\log_2 N) + 2(\log_2 d) + 3(\log_2 k) = \cancel{8\log_2(16) + 2\log_2(8) + 3(\log_2 2)} \; \cancel{= 38+16}$$

$$Cost(Tree2, data) = 4(\log_2 N) + 4(\log_2 d) + 5(\log_2 k) \; \cancel{= 4\log_2 16 + 4\log_2 3}$$

~~Cost(Tree 1) is lesser than Cost(Tree 2)~~

~~Tree 1 is better than Tree 2~~

$$Cost(Tree 1) = 8\log_2 N + 2\log_2(3) + 3\log_2(2) = 8\log_2 N + 6.169$$

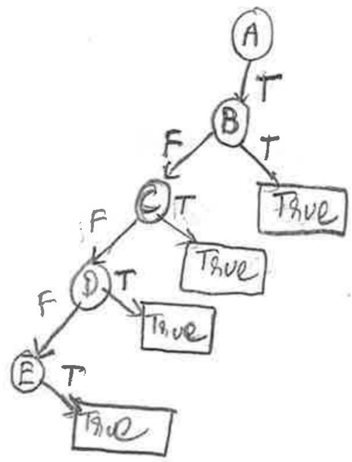
$$Cost(Tree 2) = 4\log_2 N + 4\log_2(3) + 5\log_2(2) = 4\log_2 N + 11.33$$

Cost (Tree 1) is lesser than cost (Tree 2)

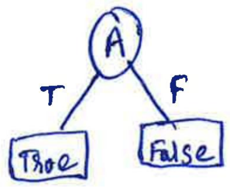
Hence Tree 1 is better.

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.

$$l(\theta|x) = p(X|\theta) = \prod_{t=1}^{N} p(x^t|\theta)$$

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

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

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

Partially diff. wrt p_g & set it to zero,

$$\frac{\partial L}{\partial p_g} = \frac{\partial}{\partial p_g}\left(\sum_{t=1}^{N} p_g (1-p_g)^{x-1}\right)$$

$$\frac{\partial L}{\partial p_g} = \sum_{t=1}^{N} \left[p_g \cdot (x-1) \cdot (1-p_g)^{x-2} + (1-p_g)^{x-1}(-1)\right]$$

$$\frac{\partial L}{\partial p_g} = \frac{\partial L}{\partial p_g} \sum_{t=1}^{N} \frac{1}{p_g} + \frac{x^t - N}{1-p_g}(-1)$$

$$\frac{\partial L}{\partial p_g} = \frac{1}{p_g} + \sum_{t=1}^{N} \frac{x^t - 1}{1-p_g} = 0$$

$$\frac{1}{p_g} + \frac{\sum_{t=1}^{N} x^t - N}{1-p_g}$$

$$1 - p_g + \sum_{t=1}^{N} x^t p_g - N p_g = 0$$

$$\sum x^t + p_g N = 0$$

$$\boxed{\hat{p}_g = -\frac{\sum x^t}{N}}$$

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**.

To incorporate prior density,

$$p(\theta|x) = \frac{p(x|\theta) \cdot p(\theta)}{\int p(x|\theta') \cdot p(\theta') \, d\theta}$$, where $p(\theta)$ is prior density.

To estimate density at x:

$$\int p(x|\theta) \; p(\theta|x) \, d\theta$$

$(x = \{x^t\})$

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?

let outlier = 20 then X = {0, 1, 2, 10, 11, 12, 20}

If we choose centroids as 1 & 10 then 2 clusters will be

C_1 = {0, 1, 2, 10} & C_2 = {11, 12, 20}. Even after centroids, clusters will be same. Hence K-mean clustering is not robust to outliers.

The closest an ~~other~~ outlier can be to other pts. & still be its cluster, is if the outlier has value of 15 or less.

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.

4. (12 points) Dimension Reduction

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

PCA is unsupervised method to reduce dimension of data by projecting it to lower dimension. $z = w^T x$ (projection of x on w). Choosing PCA with largest eigen value.

$$PoV = \frac{\lambda + \lambda_2 + \dots \lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_k + \dots \lambda_d} > 0.9$$

(λ is variance along dimension. The i^{th} principal component is the eigenvector of the i^{th} largest eigen value)

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

The similarities between PCA & LDA is that in both methods, it finds a low dimensional space such that when x is projected classes are well seperated & both are unsupervised learning.

PCA maximizes variance.

LDA maximizes fischer's linear discriminant.

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

Isomap uses local geodesic distance & Laplacian eigenmap uses similarity metric B_{rs} between r & s instances. These both metrics preserve distance among data points when projecting to lower dimension.

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

$P(\text{Yes}) = \frac{\text{Yes}}{\text{Total}} = \frac{3}{8}$

$P(\text{No}) = \frac{5}{8}$

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

$P(\text{Credit} \mid \text{No}) = \frac{2}{8}$

$P(\text{Card} \mid \text{Yes}) = 1/3$

$P(\text{Card} \mid \text{No}) = \frac{3}{8}$

~~Naive~~ bayes $P(\text{Deal} \mid \text{Yes}) = \frac{1}{3}$ $P(\text{Deal} \mid \text{No}) = \frac{2}{8}$

$P(x_i \mid C_i) = P(x_1 \mid C_i) * P(x_2 \mid C_i) * P(x_3 \mid C_i)$ for x_1 to x_m.

$P(\text{Credit, Card, Deal}) = \{P(\text{Credit} \mid \text{Yes}) * P(\text{Card} \mid \text{Yes}) * P(\text{Deal} \mid \text{Yes})\} \times P(\text{Yes})$

$\cancel{= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times = \frac{1}{27}}$

$= \frac{1}{3} \times \frac{1}{3} \times \frac{1}{3} \times \frac{3}{8} = \frac{1}{72} = 0.0138$

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

P (Credit, Card, Promotion) = P[Credit | Yes] × P [Card | Yes] × P[Promotion | Yes] × P(Yes)

$= \frac{1}{3} \times \frac{1}{3} \times \frac{3}{8} = \frac{1}{24}$

P[Promotion | Yes] is not considered by Naive Bayes as it is not there in the dataset.

Unknown terms can be handled using Laplace estimate given by $\frac{\mu}{\text{No. of classes}}$

6. (14 points) Association rules

Use the dataset in Question 4. Given the association rule: **Interest → Card**

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

$$\text{Support(Interest} \to \text{Card)} = \frac{\text{No. of total interest \& card}}{\text{No of total } \cancel{\text{only card}}} = \frac{3}{8} = 0.375$$

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

$$\text{Confidence(Interest} \to \text{Card)} = \frac{\text{No. of total with interest \& card}}{\text{No. of total with interest}} = \frac{3}{4} = 0.75$$

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

Credit Interest → Card

Then this rule can be pruned:

Interest → Card Credit

According to Apriori algorithm, For (x, y, z) to be frequent, (x, y), (x, z), (y, z) should be frequent. If one not frequent, none supersets can be frequent.

Consider superset { Credit, Interest, Card },

Given that Confidence { Credit Interest → Card } is low means that confidence of its subsets (Interest → Card, Credit) & (Credit → Card, Interest) must also be low, hence cannot be pruned.

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>