20231_dsci_552_32416 Midterm

Kayvan Shah

TOTAL POINTS

55.5 / 110

QUESTION 1

Decision Trees and Bias/Variance Dilemma 42 pts

1.1 **3 / 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 *Point adjustment*

Wrong explanation of bias and variance

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

- 0 pts Correct

- 2 *Point adjustment*

Correct explanation. Limits for the point to be an outlier not mentioned. Negative coordinates not considered

3.2 **6 / 6**

- 0 pts Correct

+ 1 *Point adjustment*

Good job

QUESTION 4

Dimension Reduction 12 pts

4.1 **0 / 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 **3.5 / 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

It uses Gaussian Kernel constructed from neighbours within a given distance which may be calculated using euclidian Distance

QUESTION 5

Naive Bayes Classification 12 pts

5.1 **2 / 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 **0 / 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 **0 / 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: KAYVAN SHAH
Reg No: 1106-6506-85

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

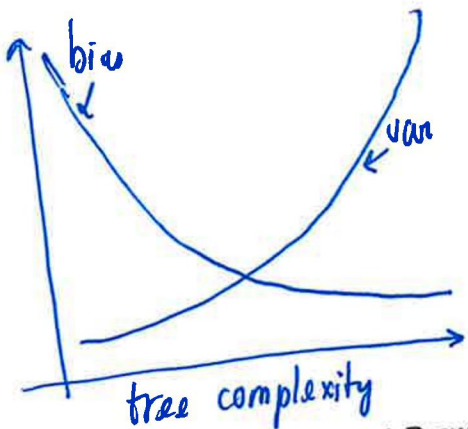
   a. (6 points) Explain the bias/variance dilemma **specifically** in the context of **decision trees**. Draw a diagram of bias/variance to illustrate your explanation. Be sure to carefully label each part of your diagram.

~~Suppose there are N attributes and M classes. Let N=4, M=3~~

Case 1

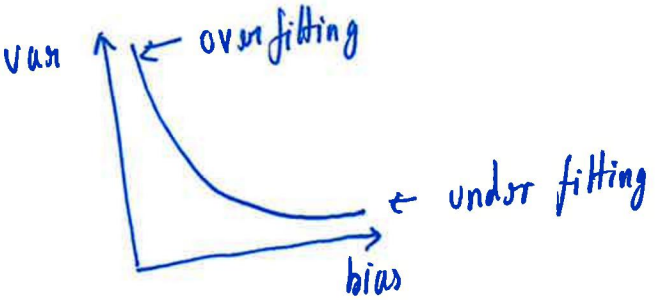
C_1
C_2
C_3

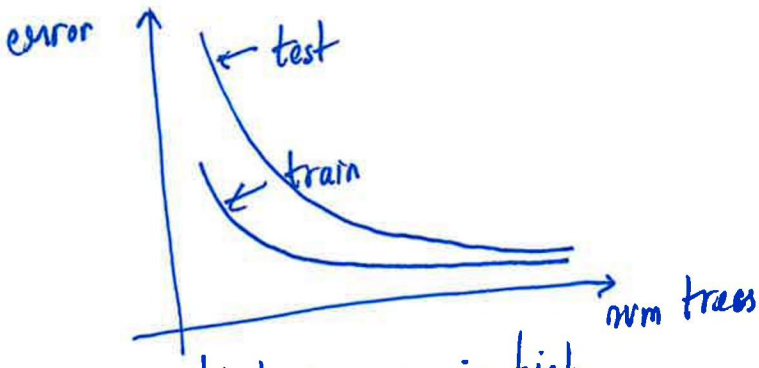
If the data is not balanced or stratified with most data points belonging to Class C_1. The decision is said to be biased towards that class. If there more number of attributes in dataset, the data distribution is sparse in the space increasing the variance and decreasing the bias.



We choose subtree with ~~mos~~ best homogenvity so that it is not baised towards one class while it should also not branch on all attr present spiking up variance.

   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?





underfitting → train error, test error are high, i.e var is high

Overfitting → train error can be low as the data is highly biased but the test error will be high as it will always be classified C_i that DT is biased too.

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.

Purpose of tree pruning is get rid redundant nodes and branches from the tree yielding high error. Doing this we can ~~at~~ reduce the effect of attr ~~error on~~ that are non significant on the model performance.

Types of tree pruning:

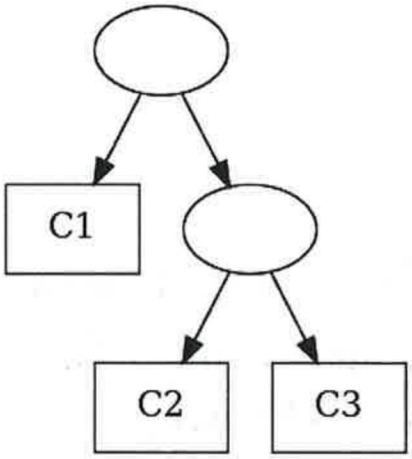
1) ~~Forward~~ Pre pruning → stop early (spliting and branching); i.e when we find no significant change in error

2) ~~Backward~~ post pruning → exhaust all attr and cases by spliting and branching and then check for error, while pruning those nodes and branches having the most error.

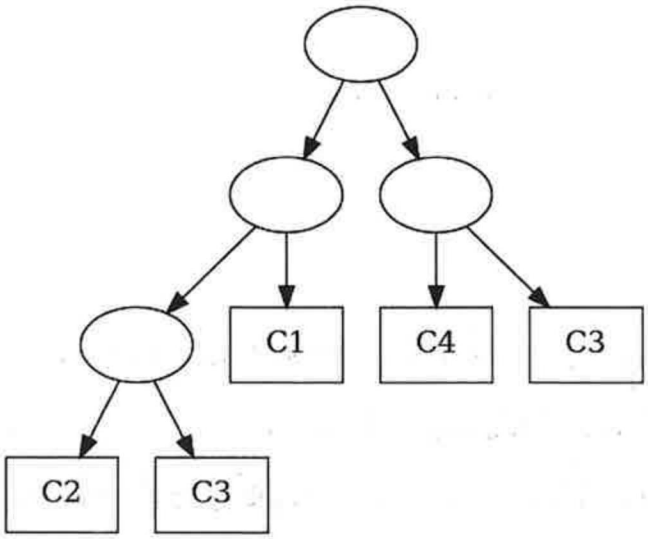
e. (8 points) Minimum description length (MDL) principle.
Consider the two decision trees below. Assume they are generated from a dataset of 16 binary attributes and 4 classes, C_1, C_2, C_3 and C_4. Assume

- 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 is better

cost (model) is low

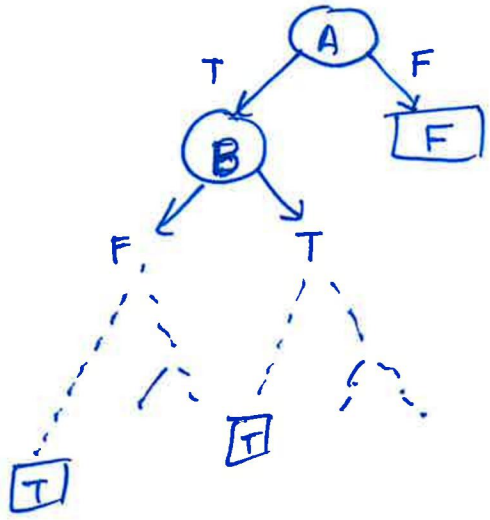
cost (~~data/model~~ model/data) is low

hence MDL is low

has least splits and nodes.

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).

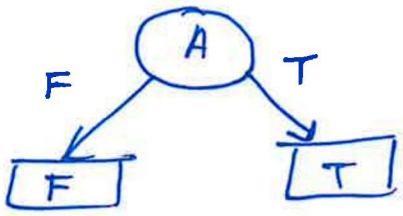
A	B	C	D	E	Class
T	F	F	F	F	T
T	T	F	F	F	T
T	F	T	F	F	T
T	F	F	T	F	T
T	F	F	F	T	T
F	T	T	T	T	F
F	F	F	F	F	F
F	T	F	F	F	F
F	F	T	F	F	F
F	F	F	T	F	F
F	F	F	F	T	F



If A is False the output will always be F and vice versa regardless of values for B,C,D,E

if A is ~~true~~ T Class is T

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.

$$\mathcal{L}(p_g|X) = \log \prod_{t=1}^{N} \left[(1-p_g)^{x^t-1} \cdot p_g\right] \leftarrow \text{likelihood}$$

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

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

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

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

$$= \sum_t (x^t - 1)\log(1-p_g) + \sum_t \log p_g$$

$$= \sum_t x^t \log(1-p_g) - \sum_t \log(1-p_g) + \sum_t \log p_g$$

$$\text{MLE}_{p_g} = \arg\max_{p_g} \{\mathcal{L}(p_g|X)\} = p_g^*$$

$$\frac{\partial \mathcal{L}}{\partial p_g} = 0 = \sum_t x^t \cdot \frac{1}{1-p_g} - \sum_t \frac{1}{1-p_g} + \sum_t \frac{1}{p_g}$$

$$= \frac{1}{1-p_g}\sum x^t - \frac{1}{1-p_g}\sum_t 1 + \frac{1}{p_g}\sum_t 1 = \frac{\sum_t (x^t - 1)}{1-p_g} + \frac{N}{p_g}$$

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

$$p_g \sum_t x^t - N p_g + N - N p_g = 0 \leftarrow p_g \sum_t x^t - p_g(N) + (1-p_g)N = 0$$

$$p_g(1-p_g)$$

$$-p_g \sum_t x^t + 2N p_g = N$$

$$p_g\left(2N - \sum_t x^t\right) = N \Rightarrow p_g = \frac{N}{2N - \sum_t x^t}$$

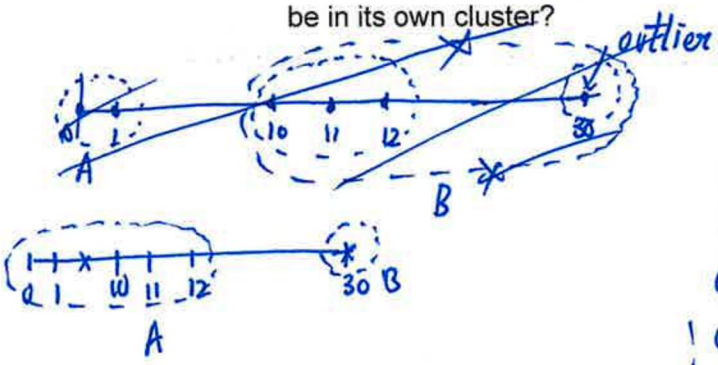
c. (4 points) To the surprise of the entomologist the beetles in this dataset only needed about half the number of attempts as reported in entomology literature. Suppose the entomologist was able to obtain the prior density from literature. Write down the equation the entomologist needs to solve to incorporate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

$\theta_{MAP} = \arg\max_{p_g} p(p_g|X)\, p(X|p_g)$

here we maximize the posteriori

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 the outlier is far enough from the centroids of other clusters, K-means will choose to form a different cluster for it which ~~com~~ is cluster B. since the distance from centriod is greater than that of cluster A.

centroid for A lies between pts in X while the centroid lies on outlier itself for B.

For outlier to be in its own cluster the point has to far enagh from ~~the~~ A. avg distance of points in A from centroid has to large enough such that its not in any proximity to pts in A.

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.

If $d_k(x)$ for instance is less than $ave_k(x)$ than that instance belongs to that cluster while for an instance whose $d_k(x) > ave_k(x)$, a new ~~may be~~ will be formed for it and if $d_k(x) >> ave_k(x)$ that instance is considered as an outlier wrt all the clusters present in the space.

For above ex: $d_B(30) > ~~out~~_A\ avg_A(30)$ hence it an outlier, ∴ separate cluster while $d_A(\{1,2\}) > avg_A(\{1,2\})$ ∴ forming one single cluster.

4. (12 points) Dimension Reduction

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

eigenvalue λ_i of the i^{th} component represent the projection of i^{th} component of the attr vector on the lower dimension

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

Both them uses dot product to reduce the dimension of data.

PCA can be used for both classification and regression. LDA is specific to classification tasks

PCA is sensitive to outliers while LDA is not so sensitive as PCA.

Both of the methods are linear.

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

ISOMAP uses geodesic curve distance that computes distance blw all the points

Laplacian Eigenmaps uses Euclidean distance which computes distance blw neighbouring points.

Both of those metric maintains the pairwise distance matrix when reduce down lower dimension from higher 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

$P(\checkmark) = 3/8$

$P(X) = 5/8$

Credit → card → deal

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

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) = #{instance with both interest and card} / #{instances}

= 3/8

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

Confidence(interest → card) = P(card | interest) = #{instances with both card, interest} / #{instances with card}

= 3/5

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) = 3/3

#{inst with cred, int, card} / #{inst card} = 1/5

Confidence(interest → card credit) = #{card credit int} / #{int} = 1/4

since both the rules have low confidence they can be pruned.

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

<extra sheet>

$p(x) = (1-p_g)^{x-1} p_g$

$\prod_{t=0}^{N} (1-p_g)^{x^t-1} p_g$

$\prod_{t=0}^{N} (1-p_g)^{x^t-1} p_g$

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

$p(x)$

$\prod_{t=0}^{N} \left(1 - \frac{1}{x^t}\right)^{x^t-1} \frac{1}{x^t}$

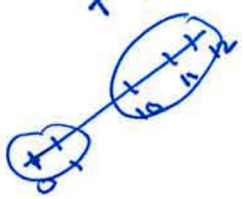
$\left(\frac{x^t-1}{x^t}\right)^{x^t-1} \cdot \frac{1}{x^t}$

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

$\sum (x^t - 1) \log\left(\frac{x^t-1}{x^t}\right) + \log\left(\frac{1}{x^t}\right)$

$(x^t - 1)\log(x^t - 1) - (x^t + 1)\log x^t$

$+ \log(A) - \log(x^t)$



<extra sheet>