CSE 4/587 - Midterm Exam #1

ANSWER KEY

Name: _____

UBIT: _____

Person #: _____

Seat #: _____

Academic Integrity

My signature on this cover sheet indicates that I agree to abide by the academic integrity policies of this course, the department, and university, and that this exam is my own work.

Signature: _____ Date: _____

Instructions

1. This exam contains **3 double-sided pages, plus this cover sheet and a scrap sheet in the back**. Be sure you have all the pages before you begin.
2. Clearly write your name, UBIT name, person number, and seat number above. **Additionally, write your UBIT name at the top of every page now.**
3. You have 50 minutes to complete this exam. Show all work where appropriate, but **keep your answers concise and to the point.**
4. After completing the exam, sign the academic integrity statement above. Be prepared to present your UB card upon submission of the exam paper.
5. You must turn in all of your work. No part of this exam booklet may leave the classroom.

DO NOT WRITE BELOW

Q1	Q2	Q3	Total
25	25	25	75

Question 1 - DIC Overview [25 Points]

a) Name the four Vs of Big Data [4 points]

Volume, Veracity, Variety, Velocity

[1 point per correct answer]

b) In one sentence, describe the goal of: [6 points]

i) Data cleaning

Any answer that demonstrates understanding that the goa of data cleaning is to prepare the data for further analysis

[3 points for correct answer, 2 points if they just give examples]

ii) EDA

Any answer that demonstrates understanding that the goal of EDA is to learn about the nature of the data

[3 points for correct answer, 2 points if they just give examples]

c) Name 3 distinct EDA operations. For each one describe in one sentence an example of what you might learn from the operation [9 points]

Possible answers:

1. Creating histograms to learn about the distribution of the data
2. Scatter plots to learn about the relation between two features
3. Box and whisker plot to see how a feature varies w.r.t. another one
4. Determining min/max/mean to learn about the range of the data, find outliers, etc
5. Get a description of features/types to learn about what data you have
6. Enumerate missing/nan values to understand how much data is missing
7. Any other reasonable operation that will teach you about the nature of the data

[1 point per operation, 2 points per valid explanation (can award partial for explanation)]

d) For each of the following, determine whether you would use an algorithm to **predict**, **cluster**, or **classify.** [6 points]

i) Determining the sale price of a stock given historical prices

Predict

ii) Categorize forum users into distinct categories based on usage habits

Cluster

iii) Determine whether or not it will rain on a given day

Classify

iv) Fill in missing artists in your music collection

Classify

v) Estimate how many people will show up to a concert

Predict

vi) Divide your dataset in order to apply different models to each segment

Cluster

[1 point per each correct answer]

Question 2 - Models and Algorithms [25 Points]

a) Name two characteristics that distinguish k-Means from the other algorithms that we have covered in class. [4 points]

Any of the following:

1. Unsupervised
2. Hard to interpret results
3. Often used to preprocess
4. Clustering instead of prediction or classification
5. Other reasonable answer

[2 points per correct answer (can award partial credit)]

b) Name the 3 classifiers we have covered in class. For each one and state whether it is binary or multiclass, and whether it is structural or statistical. [9 points]

k-NN: structural mutliclass
Naive Bayes: statistical binary
Logistic Regression: statistical binary

[1 point per correct classifier, 1 per correct structural vs statistical, 1 per binary vs multi]

c) Explain the difference between statistical models and machine learning algorithms in terms of how they relate the data to the real world. [2 points]

Statistical models attempt to model the process. ML attempt to find patterns in the data.

[2 points for a correct answer]

d) In linear regression, explain what the error term captures and why we need to use it in our linear regression models. [4 points]

Error term captures the variance that is not described by our model. We use it because there is always uncertainty/incompleteness in the data, so our model will never be perfect. Alternatively, we use it to directly model the variance in our prediction.

[2 points for what it captures, 2 points for why we use it]

e) From the following classification data, compute accuracy, precision, and recall [6 points]

	Actually Positive	Actually Negative
Predicted Positive	5	30
Predicted Negative	20	45

Accuracy: 50 / 100 = 50%
Precision: 5/35 = 14%
Recall: 5/25 = 20%

[2 points per correct answer]

Question 3 - Naive Bayes [25 Points]

Imagine we are trying to predict the weather based on the hat-wearing habits of the UB student population. Specifically, we have conducted a study over the past 4 academic years with 1000 student volunteers. Each day we have collected data about whether it was sunny or not, and which student volunteers wore hats that day.

The dataset contains information for 1200 days
180 of these days were categorized as sunny
Each day also has recorded whether or not each of the 1000 volunteers wore a hat

a) What are the "features" of this dataset, and how many features are there? [5 points]

The features are the student volunteers. There are 1000.

[4 points for understanding the students are features, 1 point for 1000]

b) Based on our analysis, Alice wore a hat on 400 days over the duration of the experiment. 150 of these days were sunny days. Given these counts, use Bayes Law to estimate the probability that Alice is wearing a hat on a sunny day. [5 points]

p(Alice | sunny) = 150 / 180 = 0.83

[5 points for correct answer]
[4 points if minor math error]
[2 points if correct formula but wrong probability computed for one of the terms]
[2 points if they correctly compute p(sunny | Alice) instead]

c) We've done similar analysis on Bob, Carly, Devon, and Eustace. The probability that each of these students wears a hat on a sunny day is 0.75, 0.20, 0.35, and 0.90 respectively. One day, you observe all 5 of them in class, and Alice, Carly, and Eustace are wearing hats. Bob and Devon are not. You know it is sunny out today. What is the probability of observing this configuration of hats? [5 points]

p(x | sunny) = p(Alice | sunny) * p(Carly | sunny) * p(Eustace | sunny) * (1 - p(Bob | sunny)) * 1 - p(Devon | sunny))
= 0.83 * 0.20 * 0.90 * 0.25 * 0.65 = 0.02

[5 points for correct answer, 4 points for minor math error]
[2 points for correctly computing answer without Alice, or with wrong number for Alice]

d) Later that week, you observe the same situation, but cannot remember if it is sunny or not. You know that in general, the odds of observing this exact configuration of hats is p(x) = 0.005. What is the probability that it is sunny out? [5 points]

p(sunny | x) = p(x | sunny) * p(sunny) / p(x) = 0.02 * 180/1200 / 0.005 = 0.6

[5 points for correct answer, 4 points for minor math error]

e) When looking through the dataset, you notice that Felicia didn't wear a hat on any of the 180 sunny days. If you see her wearing a hat, what does our model say about the probability that it is sunny out? Name two ways we could address this potential problem. [5 points]

Model says 0% chance that it is sunny out. Fix with more data, or laplace.

[1 point for knowing the model will say 0%, 2 points per correct fix]

Scrap Paper