# CSE 4/587 - Practice Midterm Exam #1

Below you will find questions taken from the FA22 midterm and final exams that relate to the content we've covered in SP23 thus far. These questions should be used to familiarize yourself with the styles of questions asked in this class. **It is by no means an exhaustive list of content that may show up on the midterm, nor is it indicative of the length of the midterm.** Additionally, some questions may contain content that has not been covered yet in this semester's offering of the course.

Name: ______________________________ Person #: ________________________

UBIT: ______________________________ 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 9 total pages (including this cover sheet). 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 | Q4 | Total |
|---|---|---|---|---|
| | | | | |
| 20 | 15 | 35 | 10 | 70 |

## Question 1 - Modeling and Algorithms [20 Points]

a. How can we help ensure that our supervised learning models are not overfit? [4 points]

**[4 points] Split data into training and test. Fit against training, compare against test. Full credit for anything related to the above answer, or other answers that are correct. Partial (2 points) for mentioning training set but not saying how it is used.**

b. In linear regression, what does the error term, ε, capture? What about $R^2$? [4 points]

**[2 points] E captures [variance, noise, effects from features not included in our model]**

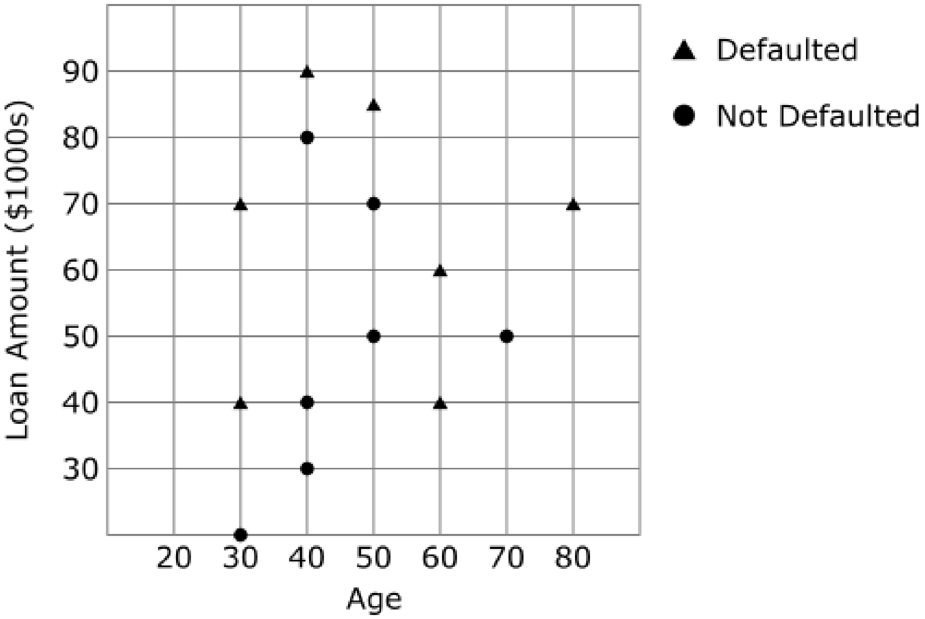
**[2 points] R2 captures [the effectiveness/accuracy of our model, the amount of the variance our model captures]**

**Give 1 point for either of the above if they give just a formula for how to calculate them but don't explain what they mean.**

c. Name the four parameters we need to define in order to fully specify and evaluate a K-NN model for a given dataset. [4 points]

**[1 point] Similarity / distance metric to define how similar / close two points are**
**[1 point] scaling done on the features**
**[1 point] evaluation metric (ie accuracy, precision, recall, etc)**
**[1 point] k**

**Below is a plot containing a number of data points with a known classification.**
The x-axis represents a person's age, and the y-axis represents a loan amount in thousands of dollars. Each point is classified as ● or ▲. Points classified as ▲ represent people that have defaulted on their loan (did not pay it back). Points classified as ● represent people who have not.



d. If we assume euclidean distance determines the similarity of two points, does K-NN predict that a 40 year old client with a loan of $70k will default for k=3? What about for k=5? [4 points]

**[2 points] k=3 predicts Not Default**
**[2 points] k=5 predicts Default**

e. Explain what would happen if we give the loan amount in terms of dollars instead of thousands of dollars. Based on that explanation, what would our model predict for the same client from part (d) with k=3? [4 points]

**[2 points] Explain that changing the scaling would result in the loan amount becoming the dominant feature in determining distance between two points (or some similar answer explaining that the scaling will change the results)**

**[2 points] With the rescaled data, k=3 now predicts Default**

## Question 2 - Classifiers [15 Points]

a) Name the three classification algorithms we described in lecture. For each algorithm state whether it is a structural classifier, or statistical classifier. [6 points]

**1 point** per classifier mentioned [k-NN, Naive Bayes, Logistic Regression]
**1 point** per correct categorization [structural, statistical, statistical]

b) In the second half of the class we used two motivating examples when discussing classifiers: spam email classification, and ad click probability. State what was used as the features for each example, and give one reason why k-NN may not be effective for these two use cases. [4 points]

**1 point** for stating that words are the features in spam classification
**1 point** for stating that urls visited are the features in ad click probability
**2 points** for stating that k-NN works better for low dimensionality (or Curse of Dim)

c) Naive bayes is an exceedingly simple algorithm. Name one reason why it can still be effective for something like spam classification. [2 points]

**2 points** for mentioning that it works well with large amounts of data, or independent features

d) How can we adapt binary classifiers to multi-class problems? [3 points]

**3 points** for mentioning we run on each class, and pick the most likely

## Question 3 - Naive Bayes [35 Points]

The faculty and staff at UB receive a lot of emails every day, but many of these emails are actually spam. We'd like to use our knowledge from CSE 4/587 to automatically detect spam emails and filter them out. In order to do so, we've already collected a large sample of emails and classified them as spam or not spam, and ran a word counting algorithm over the dataset.

The dataset contains 1000 emails
300 of these emails are categorized as spam

a) Based on our word count analysis, the word "buy" shows up in 75 emails. 70 of these emails that contain the word "buy" are categorized as spam. Given these counts, use Bayes Law to determine the probability that an email containing the word "buy" is spam. [10 points]

**10 points** for getting the correct answer (or close with just math errors)
**5 points** for having some correct steps/formulas

p(spam) = 300 spam emails / 1000 total emails = 0.3

p("buy") = 75 occurrences / 1000 total emails = 0.075

p("buy" | spam) = 70 occurrences in 300 spam emails = 70 / 300 = 0.233

p(spam | "buy") = p("buy" | spam) * p(spam) / p("buy")

= (70/300) * (300/1000) / (75 /1000) = 70/75 = 0.9333

b) Assume we have done similar computations as in part (a) to determine that the probabilities that "computer", "won", "faculty", and "meeting" show up in a spam email are 0.1, 0.85, 0.01, and 0.05 respectively. Now we have an uncategorized email that contains the words "computer", "buy", and "meeting", and does not contain "won", or "faculty". Use the principles we've discussed in class to determine the probability that this exact set of words show up in an email that is **known to be spam**. [10 points]

**10 points** for getting the correct answer (or close with just math errors)
**5 points** for having some correct steps/formulas

x = ["buy", "computer", "won", "faculty", "meeting"] = [1,1,0,0,1]

p(x | spam) = p("buy" | spam) * p("computer" | spam) * (1 - p("won" | spam)) * (1 - p("faculty" | spam)) * p("meeting" | spam)

= (70/300) * 0.1 * 0.15 * 0.99 * 0.05 = 0.00017

c) Given your answer from part (b), if we have determined that the probability that that exact combination of words shows up in any email is 0.02, what is the probability that our uncategorized email containing those words is spam? [6 points]

**6 points** for getting the correct answer (or close with just math errors)
**3 points** for having some correct steps/formulas

p(x) = 0.02

p(spam | x) = p(x | spam) * p(spam) / p(x) = 0.00017 * 0.3 / 0.02 = 0.00255

d) Let's assume that based on our word count analysis, the word "professor" does not show up in any spam emails in our dataset. This would result in the probability that an email is spam if it contains the word "professor" to be 0.0, which may have a negative impact on our implementation of Naive Bayes. Name **two** ways we could address this potential problem. [4 points]

**2 points** for mentioning Laplace Smoothing
**2 points** for mentioning getting more data (or other reasonable answer)

e) We've deployed our spam filter but now there seems to be a problem… We've received some complaints from faculty that they are still receiving almost as many spam emails as they were before. When we look into these complaints, we realize that these complaints are mostly coming from female faculty members. Note that female faculty make up around 16% of the total faculty in the CSE department. Give one possible explanation as to why our model is more effective at filtering out spam targeted at male faculty but not those targeting female faculty. [5 points]

**N/A for SP23 Midterm #1**

## Question 4 - Misc [10 Points]

a) What are the four Vs of data intensive computing? [2 points]

**0.5 points** (rounded down) per correct answer [volume, velocity, variety, veracity]

b) Name two sources of uncertainty in data. [2 points]

**1 point** per correct answer [random process generating data, random sampling]

c) Briefly explain why fault tolerance in distributed systems is important. [2 points]

**N/A for Sp23 Midterm #1**

d) Name one concrete example of a data intensive application that you interact with on a frequent basis. [2 points]

**2 points** for any feasible answer (Netflix, Social media, specific recommendation engines, Google search, etc)

e) Name two solutions you could employ if the dataset you are working on becomes too big to store/process in a timely manner. [2 points]

**1 point** for mentioning sampling
Rest is **N/A** for Sp23 Midterm #1

Scrap Paper