Practice Midterm

Please read all the guidelines carefully before starting the midterm. There are **eight (8)** questions. Each question is worth **10 points**.

Due date: March 16, 3:30 PM.

Grading

Your answers will be evaluated based on the following criteria:

- **Completeness:** Your answers will be checked for completeness. Specifically, for a question that requires several steps of thinking / writing / calculation, we expect you to complete the full range of steps to answer the question. The range of steps that needs to be completed will be determined by the course material and the specific nature of the question.
- **Correctness:** Your answers will be checked for correctness. For your answer to be correct, you need to have the correct answer, correct implementation, and the correct result. 'Correct' means that you follow the steps suggested by the assignment and your instructor and obtain the expected result without making any theoretical / mathematical mistakes. 'Correctness' is not a binary term, there may be varying degrees of correctness; and, your grade will be evaluated based on how different your answer is from the expected result.
- **Format:** An indispensable part of every assignment is the format. To make sure that your assignment can be read and processed easily, we expect you to follow the guidelines set by the instructor. These guidelines may include specific requirements about text-based answers, mathematical calculations etc.
- **Academic Honesty:** Your submission must fulfill the academic honesty expectations set by the instructor and put forward in the syllabus. Specifically, when expected, you need to produce work within the limits defined in the syllabus – some of the assignments may require you to work individually, and some others in a team. For more information, please read the syllabus.

Midterm Guidelines – Before You Start

1) **You must complete the midterm on your own.**
2) **You cannot use your laptop, your smartphone, your calculator, your watch, your smart glasses, and other smart devices.**
3) **You cannot use your notes or your textbook or any other written material.**
4) **Please answer only one (1) question per sheet.**
5) **Please write your name on each sheet. If we cannot find your name, you won't be graded and will lose your points for that question.**
6) **Make sure that you explain concepts carefully and in full detail. Make sure you also use examples when necessary.**
7) **The exam will end promptly on due date. Please calculate accordingly and finish on time.**

Deliverables:

1) Eight (8) sheets each containing an answer to a question. **One question per sheet**. All sheets need to have **your first name and last name**.

Questions

For Q1, Q2 and Q3, to receive full points, you will need to show all the steps in your calculation.

1) **[10 points]** Please answer the following questions about matrix algebra:
   a. **[5 points]** Calculate the inverse of the following matrix: A = $\begin{pmatrix} 3 & 5 \\ -7 & 2 \end{pmatrix}$.
   b. **[5 points]** Show that *AA*$^{-1}$ *= Identity Matrix* (show all the steps in your calculation).

2) **[10 points]** Calculate the *cosine similarity* between the following two sentences:
   Sentence 1: "I like hamburger more than fried chicken."
   Sentence 2: "I like sushi more than hamburger."

   Formula for *cosine similarity*: $\frac{A*B}{|A|*|B|}$, where *A* and *B* are the vectorized forms of sentences.

   a. **[5 points]** Vectorize the sentences above and write the vectorized forms. (*Hint*: Apply count vectorization)
   b. **[5 points]** Calculate the *cosine similarity* by applying the formula above (reporting the ratio will be fine – no need to calculate decimal values).

3) **[10 points]** You are a person who lives in Rochester and you would like to go on a picnic in the upcoming Spring. Rochester is a cloudy city – specifically, 60% of all days are cloudy [on average throughout the year]. In the Spring, 30% of the days are rainy. And, from experience, we know that 50% of all rainy days start off cloudy. As a scientific mind, you don't think it is worth going on a picnic, if the probability of having rain is more than 45% on a given day. If you are interested in going on a picnic and if there were no clouds in the morning, would you go on a picnic? Using the ***Bayes' theorem***, calculate the chance of having rain, and answer the following question: Should you go on a picnic?

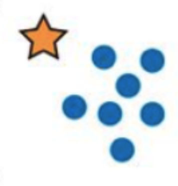
4) **[10 points]** Explain the difference between ***generative*** and ***discriminative*** models, specifically:

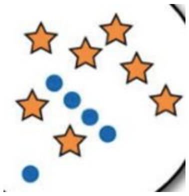
   a. **[3 points]** Provide the definition / philosophy for ***generative*** models, and give at least one example from algorithms.
   b. **[3 points]** Provide the definition / philosophy for ***discriminative*** models, and give at least one example from algorithms.
   c. **[4 points]** Let's say we would like to classify different types of Iris flower species. How would we do it using a generative model? (i) How would we do it using a discriminative model? (ii). Explain the difference by providing examples and showing your logic (hypothetical examples are fine, you need to elaborate on the mathematical differences).

5) **[10 points]** What are the four (4) different types of clustering families? (Explain each family with one or two sentences, giving examples will help) (**2.5 points** each).

6) **[10 points]** Explain how ***kmeans++*** algorithm works step-by-step by including all details in each step.

7) **[10 points]** Take a look at the datasets below, and answer the following about entropy:





a. **[5 points]** Which dataset of stars and circles have a higher ***entropy*** value?

b. **[5 points]** Calculate the entropy value for each dataset (just plugging the values into the formula will be fine – no need to calculate the actual results).

8) **[10 points]** One way of comparing two `strings` is to compute the ***Hamming distance*** between them. Thus, ***Hamming distance*** provides a way of calculating the edit distance between two `strings`. Two `strings` are different from each other if at least some characters are different. Here are a few examples that show the ***Hamming distance*** between two `strings`. Red color indicates different characters.

<u>***Hamming distance*** between:</u>

- `"karolin" and "kathrin" is 3.`
- `"karolin" and "kerstin" is 3.`
- `"kathrin" and "kerstin" is 4.`
- `0000 and 1111 is 4.`
- `2173896 and 2233796 is 3.`

a. **[8 points]** Provide a generalizable `Python` code for calculating ***Hamming distance*** [*Hint*: The code shouldn't take more than 10 lines – and there is a 1-line `Pythonic` version of it.]

b. **[2 points]** Provide at least <u>one (1)</u> example/argument that demonstrates a disadvantage of using ***Hamming distance***.