CSCE 421 :: Machine Learning :: Texas A&M University :: Fall 2021

Written Assignment 1 (W1)

Due on Wednesday, Sep 15, 11:59 PM. **Total points: 100**

Name:
UIN:

Instructions:

- This PDF has blank spaces left out after each question for you to fill in your solutions.
- You are free to either use LaTeX(use the provided *.tex* file) or handwrite them as long as they are legible.
- Refer to the submission instructions on the course webpage to upload your solutions to Canvas.
- **SHOW YOUR WORK.**

---

1 Linear Algebra and Probability Review

Question 1: Matrix Multiplication. **(10 points)**
NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this question. Show your work.

In this question you are required to perform matrix multiplication.

(1) (3 points)

$$\begin{bmatrix} 10 \\ -5 \\ 2 \\ 8 \end{bmatrix} \begin{bmatrix} 0 & 3 & 0 & 1 \end{bmatrix} = ?$$

(2) (3 points)

$$\begin{bmatrix} 7 & -3 & 1 & 9 \end{bmatrix} \begin{bmatrix} -3 \\ -4 \\ 6 \\ 0 \end{bmatrix} = ?$$

(3) (4 points)

$$\begin{bmatrix} 1 & -1 & 6 & 7 \\ 9 & 0 & 8 & 1 \\ -8 & 1 & 2 & 3 \\ 10 & 4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 6 & 2 & 0 \\ 0 & -1 & 1 \\ -3 & 0 & 4 \\ 3 & 4 & 7 \end{bmatrix} =?$$

Question 2: Vector Norms. **(8 points)**
NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this Question. Show your work.

Consider these two points in the 3-dimensional space:

$$\mathbf{a} = \begin{bmatrix} 7 \\ 0 \\ -1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} 7 \\ 9 \\ -5 \end{bmatrix}$$

Calculate their distance using the following norms.

(1) ℓ_0 (2 points)

(2) ℓ_1 (2 points)

(3) ℓ_2 (2 points)

(4) ℓ_∞ (2 points)

Question 3: Probability Calculation. **(12 points)**
NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this Question. Show your work.

Consider a situation where we are rolling 2 dices where each dice has 6 faces numbered from 1 to 6. Answer the following questions:

(1) What is the size of the sample space? (4 points)

(2) If the event we are interested in is the sum being 11, what would be the probability of observing such an event? (4 points)

(3) If the event we are interested in is the sum being 6, what would be the probability of observing such an event? (4 points)

Question 4: Mean/Variance Calculation. **(10 points)**
NOTE: This is not a programming assignment, so you may NOT use programming tools to help solve this Question. Show your work.

Assume we have a random variable X with a Uniform probability density function. Uniform probability density is defined as:

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b \\ 0 & otherwise \end{cases}$$

(1) What is the mean of X? (5 points)

(2) What is the standard deviation of X? (5 points)

2 k-Nearest Neighbors

Question 5: Feature Scaling. **(15 points)**

(1) Why is feature scaling necessary in k-NN? (5 points)

(2) What is the potential issue with using categorical variables in k-NN? (5 points)

(3) How would you pre-process categorical variables to make them suitable for use with k-NN? (5 points)

Question 6: Curse of dimensionality. **(15 points)**

(1) Briefly explain why k-NN suffers when the data dimensionality is high? (15 points)

Question 7: Choosing the best "k". (15 points)

(1) Briefly explain a strategy you could use to choose the best "k" value for a given dataset for k-NN? (15 points)

(*Hint: cross-validation.*)

Question 8: Algorithmic Complexity. (15 points)

(1) Let n be the number of training examples in a dataset and m be the dimensionality of each data point. For simplicity, assuming $n \gg m$, the algorithmic complexity for classifying a query point using naive k-NN implementation is $\mathcal{O}(k \times n)$. Can you think of a way to make it more efficient? (15 points)

(*Hint:* $\mathcal{O}(n \log(k))$)