MATH 123 Math Aspects of Data Analysis - Spring 2023
Tufts University, Department of Mathematics
Instructor: James M. Murphy
Practice Midterm 2

Instructions: This exam has four questions and is out of a total of 100 points. Each question is worth 25 points. No graphing calculators, books, or notes are allowed. Be sure to show all work for all problems. No credit will be given for answers without work shown. If you do not have enough room in the space provided you may use additional paper. Be sure to clearly label each problem and attach them to the exam. You have 75 minutes. Good luck! :-)

Your Printed Name: _____

Problem	Score
1	
2	
3	
4	
Total	

Academic Honesty Certification:

I certify that I have taken this exam without the aid of unauthorized people or objects.

Signature: _____ Date: _____

Question 1

(a) Describe, in detailed pseudocode, the DBSCAN clustering algorithm.

(b) Draw or describe in detail a data set on which DBSCAN can outperform K-means clustering. Explain.

Question 2

Let $L \in \mathbb{R}^{n \times n}$ be an (unnormalized) graph Laplacian corresponding to some weight matrix $W \in \mathbb{R}^{n \times n}$.

(a) Prove that L has at least one eigenvalue of 0.

(b) Describe in detail an example of a W matrix for which the corresponding L has exactly $k = 3$ eigenvalues of 0.

Question 3

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be data with labels $\{y_i\}_{i=1}^n,\ y_i \in \{-1, 1\}$.

(a) Explain, in detailed pseudocode, the k-NN classifier.

(b) What is the complexity of labeling a new data point with this algorithm?

(c) Should this be considered an acceptable complexity or not? Discuss.

Question 4

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be data with labels $\{y_i\}_{i=1}^n,\ y_i \in \{-1, 1\}$.

(a) Write the loss function for the hard margin support vector machine. Interpret both the function to be optimized and the constraints.

(b) Why is the idea of "maximizing the margin" relevant from the standpoint of labeling new data?