CSC 3520
Machine Learning
Florida Southern College

Assignment 4: Nearest Neighbors
Due: Monday, November 19, 2018

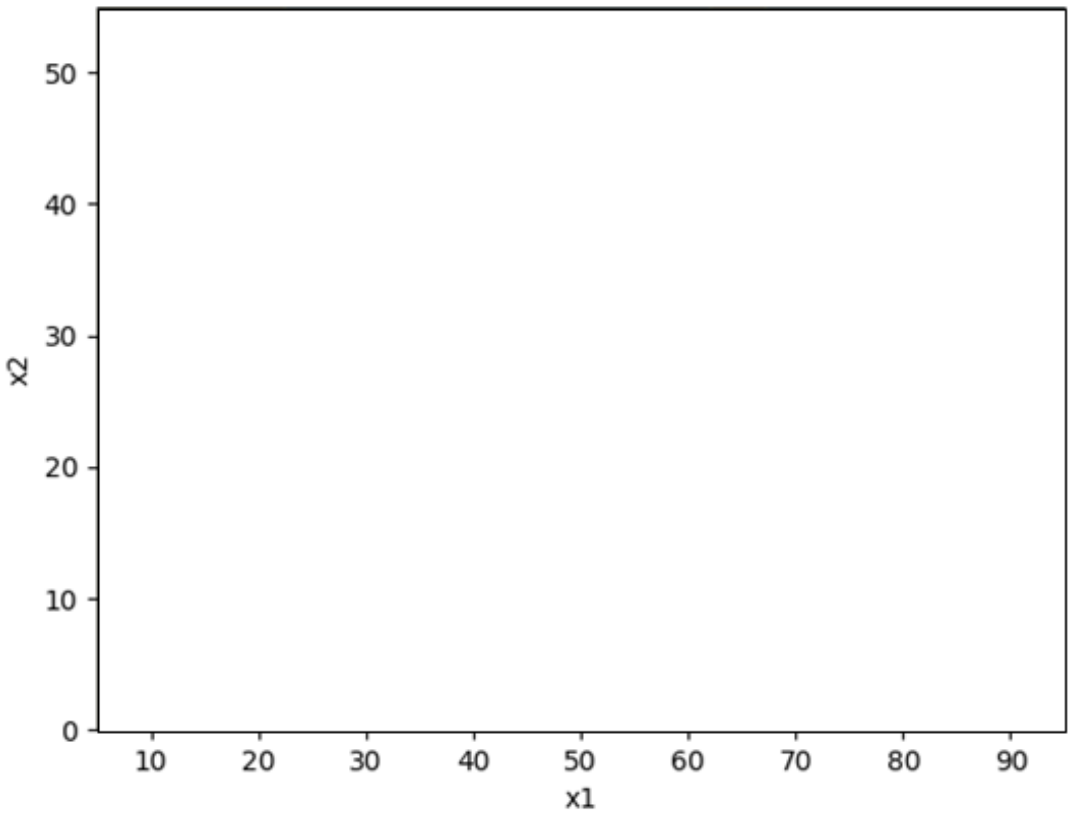
1. K-Nearest Neighbors (50 points)

Suppose you are given the follow 2D dataset:

$$X = \begin{bmatrix} 90 & 35 \\ 70 & 5 \\ 35 & 50 \\ 10 & 50 \\ 50 & 30 \end{bmatrix} \quad Y = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

where X are the features (x_1, x_2), Y are the labels, each row is a unique training sample, and there are two classes [0, 1].

(a) Plot the data (by hand) below. Use 'x' for $Y = 0$ and 'o' for $Y = 1$.



(b) On the same plot, draw the complete Voronoi diagram; that is, draw the appropriate lines based on Euclidean distance that partition the 2D space into regions such that each data point has its own region.

(c) Using a bold line, identify the decision boundary on the Voronoi diagram.

(d) Now suppose you are given three test samples:

$$X_{test} = \begin{bmatrix} 10 & 10 \\ 30 & 40 \\ 80 & 20 \end{bmatrix}$$

Add each test sample to the plot using '□'.

(e) Compute the **Manhattan distance** between each test and training sample. Fill out the table below.

		Training Data				
		1	2	3	4	5
Test Data	1					
	2					
	3					

(f) Using the distance information in the table above, determine the class label (0 or 1) for each test sample using *k*-nearest neighbors. Fill out the table below, where **each column is a different value for *k***. In the case of a tie, choose the class with the higher prior probability.

		k				
		1	2	3	4	5
Test Data	1					
	2					
	3					

2. Handwritten Digit Recognition (50 points)

Write a Python script (called handwritten_digit_recognition.py) that applies *k*-nearest neighbors to the MNIST handwritten digit dataset. Use $k = 3$. Train the classifier on all 60,000 training samples. Test the classifier on the first 100 testing samples.

(a) What is the test accuracy?

(b) Find a test sample that gets misclassified. Show the test image alongside the 3 nearest neighbors from the training set. Does it make sense why the classifier predicted the wrong digit?

(c) Try varying *k* to improve the test accuracy on the first 100 test samples. For what value of *k* can you achieve 100% accuracy?

(d) True or False: if *k*-nearest neighbors correctly predicts the digit for a given image, it is guaranteed to make the correct prediction for that same sample using (*k*+1)-nearest neighbors. Explain your reasoning in at least 2 sentences.

(e) Compare your implementation of *k*-nearest neighbors to neural networks for MNIST digit recognition. Discuss the pros and cons of each approach in at least 3-5 sentences.