Homework 1 **CSCE 633**
Due: 2.30pm on February 10, 2020

Instructions for homework submission
a) For the **math problems**, please typewrite your answers in Latex, or handwrite your solution *very clearly*. Non-visible solutions will not be graded: we wouldn't like our TA to have to guess what you are writing :)
b) For the **experimental problems**, please write a brief report. At the end of the report, please include your code. Print the report, including the code.
c) **Staple all your answers and hand them out in paper in class or during office hours.**
d) Please start early :)
e) The maximum grade for this homework, excluding bonus questions, is **10 points** (out of 100 total for the class). There are **2 bonus points**.

Question 1 (4 points)
Linear Perceptron Algorithm: The goal of this problem is to run a linear perceptron algorithm *on paper and pencil*. Assume that you have three training samples in the 2D space:

1. Sample $\mathbf{x_1}$ with coordinates $(1, 3)$ belonging to Class 1 ($y_1 = 1$)
2. Sample $\mathbf{x_2}$ with coordinates $(3, 2)$ belonging to Class 2 ($y_2 = -1$)
3. Sample $\mathbf{x_3}$ with coordinates $(4, 1)$ belonging to Class 2 ($y_2 = -1$)

The linear perceptron is initialized with a line with corresponding weight $\mathbf{w(0)} = [2, -1, 1]^T$, or else the line $2 - x + y = 0$.
In contrast to the example that we have done in class, in this problem **we will include the intercept term w_0.**

(0.5 points) (i) Plot $\mathbf{x_1}$, $\mathbf{x_2}$, and $\mathbf{x_3}$ in the given 2D space. Plot the line corresponding to weight $\mathbf{w(0)}$, as well as the direction of the weight $\mathbf{w(0)}$ on the line.

(1 point) (ii) Using the rule $sign(\mathbf{w(t)}^T\mathbf{x_n})$, please indicate the class in which samples $\mathbf{x_1}$, $\mathbf{x_2}$, and $\mathbf{x_3}$ are classified using the weight $\mathbf{w(0)}$. Which samples are not correctly classified based on this rule?
Note: You have to compute the inner product $\mathbf{w(0)}^T\mathbf{x_n}$, $n = 1, 2, 3$, and see if it is greater or less than 0.

(1.5 points) (iii) Using the weight update rule from the linear perceptron algorithm, please find the value of the new weight $\mathbf{w(1)}$ based on the misclassified sample from question **(ii)**. Find and plot the new line corresponding to weight $\mathbf{w(1)}$ in the 2D space, as well as the direction of the weight $\mathbf{w(0)}$ on the line. Indicate which samples are correctly classified and which samples are not correctly classified.
Note: The update rule is $\mathbf{w(t+1)} = \mathbf{w(t)} + y_s\mathbf{x_s}$, where $\mathbf{x_s}$ and $y_s \in \{-1, 1\}$ is the feature and class label of misclassified sample s.
Hint: The line corresponding to a vector $\mathbf{w} = [w_0, w_1, w_2]$ can be written as $w_0 + w_1x + w_2y = 0$. Make sure that you get the direction of the vector $\mathbf{w}$ correctly based on the sign of w_1 and w_2.

(1 point) (iv) Using the rule $sign(\mathbf{w(t)}^T\mathbf{x_n})$, run the linear perceptron algorithm, find and plot the weights $\mathbf{w(2)}$ and the corresponding line. Please indicate which samples are classified correctly and which samples are not classified correctly.

Question 2 (6 points)
Classifying benign vs malignant tumors: We would like to classify if a tumor is benign or malign based on its attributes. We use data from the following UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)`. Inside "Homework 1" folder on Piazza you can find three files including the train and test data (named "hw1_ question1_train.csv", "hw1_ question1_dev.csv", and "hw1_ question1_test.csv") for our experiments. The rows of these files refer to the data samples, while the columns denote the features (columns 1-9) and the class variable (column 10), as described bellow:

1. Clump Thickness: discrete values $\{1, 10\}$
2. Uniformity of Cell Size: discrete values $\{1, 10\}$
3. Uniformity of Cell Shape: discrete values $\{1, 10\}$
4. Marginal Adhesion: discrete values $\{1, 10\}$
5. Single Epithelial Cell Size: discrete values $\{1, 10\}$
6. Bare Nuclei: discrete values $\{1, 10\}$
7. Bland Chromatin: discrete values $\{1, 10\}$
8. Normal Nucleoli: discrete values $\{1, 10\}$
9. Mitoses: discrete values $\{1, 10\}$
10. Class: 2 for benign, 4 for malignant (this is the **outcome**)

(a.i) (0.5 points) Data exploration: Using the training data, compute the number of samples belonging to the benign and the number of samples belonging to the malignant case. What do you observe? Are the two classes equally distributed in the data?

(a.ii) (0.5 points) Data exploration: Using the training data, plot the histogram of each feature (i.e., 9 total histograms). How are the features distributed in the 1-10 range? Are the sample values distributed equally for each feature?

(a.iii) (1 point) Data exploration: Randomly select 5 pairs of features. Using the training data, plot scatter plots of the selected pairs (i.e., 5 total scatter plots). Use a color-coding to indicate the class in which the samples belong to (e.g., blue for benign, red for malignant). What do you observe? How separable do the data look?

(b.i) (2 points) Classification: Implement a K-Nearest Neighbor classifier (K-NN) using the euclidean distance ($l2$-norm) as a distance measure to classify between the benign and malignant classes. **Please implement K-NN and do not use available libraries.** In the report, show your code for this question.

(b.ii) (1 point) Explore different values of $K = 1, 3, 5, 7, \ldots, 19$. You will train one model for each of the ten values of K using the train data and compute the classification accuracy (Acc) and balanced classification accuracy ($BAcc$) of the model on the development set. Plot the two metrics against the different values of K. Please report the best hyper-parameter K_1 based on the Acc metric, and the best hyper-parameter K_2 based on the $BAcc$ metric. What do you observe? **Please implement this procedure from scratch and do not use available**

libraries.
Hint: $Acc = \frac{\#\ correctly\ classified\ samples}{\#\ samples}$
$BAcc = \frac{1}{2}\left(\frac{\#\ correctly\ classified\ samples\ from\ Class\ 1}{\#\ samples\ in\ Class\ 1} + \frac{\#\ correctly\ classified\ samples\ from\ Class\ 2}{\#\ samples\ in\ Class\ 2}\right)$

(b.iii) (1 point) Report the Acc and $BAcc$ metrics on the test set using K_1 and K_2. What do you observe?

(b.iv) (Bonus, 2 points) Instead of using the euclidean distance for all features, experiment with different types of distances or distance combinations, e.g. $l0$-norm or cosine similarity. Report your findings.