**Homework 1** **CSCE 633**
**Due: 11.59pm on September 26, 2021**

---

**Instructions for homework submission**
a) For the **math problems**, please typewrite your answers in Latex, or handwrite your solution *very clearly* and scan it. 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 and *include your code.*
c) Create a **single pdf** and submit it on **CANVAS**. Please do not submit .zip files or colab notebooks.
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 **3 bonus points** at the end of the homework.

**Question 1 (6 points)**
**Predicting patient post-surgery survival:** Predicting a patient's risk of death during a breast cancer surgery can help physicians to make decisions for personalized post-surgery treatment and care. We will use the Haberman's Survival Dataset, collected by the University of Chicago's Billings Hospital, to predict survival for patients who had undergone surgery for breast cancer based on attributes that have been identified as important to our problem (i.e., patient age, year of surgical operation, number of detected positive axillary lymph nodes). We will use available data from the following UCI Machine Learning Repository:
`https://archive.ics.uci.edu/ml/datasets/Haberman%27s+Survival`.

Inside "Homework 1" on CANVAS you can find three files including the train and test data (named "data_train.csv", "data_dev.csv", and "data_test.csv") for our experiments. The rows of those files refer to the data samples, while the columns denote the features (columns 1-3) and the class variable (column 4), as described bellow:

1. Age of patient at time of operation (integer)
2. Patient's year of operation (integer, ranging between 58 and 70, which correspond to years 1958-1970)
3. Number of positive axillary lymph nodes detected (integer)
4. class: the patient survived 5 years or longer (1), the patient died within 5 years after the surgery (2)

**(a.i) (0.5 points) Data exploration:** Using the training data, compute the number of samples belonging to each class. Are the classes equally distributed?

**(a.ii) (0.5 points) Data exploration:** Using the training data, plot the histogram of each feature (i.e., 3 total histograms). How are the features distributed (e.g., unimodal, bimodal, uniform distributions)?

**(a.iii) (0.5 points) Data exploration:** Using the training data, plot scatter plots of all pairs of features (i.e., 3 total scatter plots). Use a color-coding to indicate the class in which the samples belong to (e.g., blue circle for class 1, green star for class 2). What do you observe? How separable do the classes look? Are there feature combinations for which the two classes are more separable?

**(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 three classes. Please **implement K-NN and do not use available libraries.** In the report, please show your code.

**(b.ii) (1 point)** Explore different values of $K = 1, 3, 5, 7, 9, 11, 13$. You will train one model for each of the seven values of $K$ using the train data and compute the classification accuracy ($Acc$) and balanced classification accuract ($BAcc$) of the model on the development set. Plot the $Acc$ and $BAcc$ metrics on the dev set against the different values of $K$. Please report the best hyper-parameter $K^*$ based on the $Acc$ metric and the best hyper-parameter $K^{**}$ based on the $BAcc$ metric. **Please implement this procedure, including computing the accuracy metrics, from scratch and do not use available libraries.**

*Hint:* $Acc = \frac{\text{\# correctly classified samples}}{\text{\# samples}}$

$BAcc = 0.5 \cdot \frac{\text{\# correctly classified samples from class 1}}{\text{\# samples from class 1}} + 0.5 \cdot \frac{\text{\# correctly classified samples from class 2}}{\text{\# samples from class 2}}$

**(b.iii) (0.5 points)** Report the $Acc$ and $BAcc$ metrics on the test set using $K^*$ and $K^{**}$.

**(b.iv) (0.5 points)** Instead of using the euclidean distance, experiment with the $l1$-norm (i.e., Manhattan distance) for $K = 1, 3, 5, 7$. Report your findings.

**(c) (0.5 points) ML deployment:** Assume that the Memorial Herman Hospital in Houston, TX is planning to deploy this system over the next months in order to predict patient post-surgery mortality rate. What would be your thoughts / questions / concerns regarding this?

**Question 2 (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 $w_0 + w_1 x + w_2 y = 0$. 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 3 (3 Bonus points)**

Please complete the survey provided in this link: `https://tamu.qualtrics.com/jfe/form/SV_6QLUBI7p6N9ADOG`. Please use your laptop to complete this assignment.