Assignment 2

CS 484

Submission Deadline: 03/03 11:59 pm

Problem 1: Perceptron Learning (15 marks)

The dataset lab02_dataset_1.csv has a 3-dimensional input space and a class label of *Positive* and *Negative*. For this task, you are not allowed to use any functionalities of the sklearn module.

1. Write a function my_perceptron( ) which applies perceptron algorithm on the dataset to create a linear separator. my_perceptron( ) should return a 3-dimensional weight vector which can be used to create the linear separator. Use a classification threshold of 99% i.e., my_perceptron( ) will terminate once the misclassification rate is less than 1%. (10 marks)
2. Create a 3D plot which showcases the dataset in a 3D-space alongwith the linear separator you obtained from my_perceptron( ). Use two different colors to represent the data points belonging in the two classes for ease of viewing. (5 marks)

Problem 2: Naïve Bayes Learning (25 marks)

The dataset lab02_dataset_2.xlsx contains 10,302 observations on various vehicles. You will use the observations in this dataset to train models that predict the usage of a vehicle. Your models will use the following variables:

Output Label:

- **CAR_USE**. Vehicle Usage. It has two categories, namely, *Commercial* and *Private*.

Input Features:

- **CAR_TYPE**. Vehicle Type. It has six categories, namely, *Minivan*, *Panel Truck*, *Pickup*, *SUV*, *Sports Car*, and *Van*.

- **OCCUPATION**. Occupation of Vehicle Owner. It has nine categories, namely, *Clerical*, *Home Maker*, *Doctor*, *Lawyer*, *Manager*, *Professional*, *Blue Collar*, *Student*, and *Unknown*.
- **EDUCATION**. Highest Education Level of Vehicle Owner. It has five categories namely *Below High Sc*, *High School*, *Bachelors*, *Masters*, *PhD*.

You will use only observations where there are no missing values in all the above four variables. After dropping the missing values, you will use all the 100% complete observations for training your Naïve Bayes models using sklearn. For each observation, you will calculate the predicted probabilities for CAR_USE = *Commercial* and CAR_USE = *Private*. You will classify the observation in the CAR_USE category that has the highest predicted probability. In case of ties, choose *Private* category as the output.

1. You will train a Naïve Bayes model with a Laplace smoothing of 0.01. (5 marks)
2. Output the Class counts and Probabilities P(Y_j). Also display the probability of the input variables, given each output label P($X_i|Y_j$) alongwith their counts. (5 marks)
3. Let us study a couple of fictitious persons (test cases). One person works in a *Blue Collar* occupation, has an education level of *PhD*, and owns an *SUV*. Another person works in a *Managert* occupation, has a *Below High Sc* level of education, and owns a *Sports Car*. What are the Car Usage probabilities of both these people? (5 marks)
4. Generate a histogram of the predicted probabilities of CAR_USE = *Private*. The bin width is 0.05. The vertical axis is the proportion of observations. (5 marks)
5. Finally, what is the misclassification rate of the Naïve Bayes model? (5 marks)