CSCE 4930 - Introduction to Machine Learning
Assignment 2

This is an individual assignment. Rules governing academic integrity are strongly enforced without exception.

1. Using the MSE principle, derive the coefficients (a,b) of the single-dimension linear regression model from the training data. Make sure to show all your derivation steps starting from the training sample to the final formula for both coefficients
(20 points)

2. In this problem, you will be using the following cars dataset that associates the various car features (wheels, color, type and doors) to detect its class

Color	**Type**	**Doors**	**Tires**	**Class**
Red	SUV	2	Whitewall	+
Blue	Minivan	4	Whitewall	-
Green	Car	4	Whitewall	-
Red	Minivan	4	Blackwall	-
Green	Car	2	Blackwall	+
Green	SUV	4	Blackwall	-
Blue	SUV	2	Blackwall	-
Blue	Car	2	Whitewall	+
Red	SUV	2	Blackwall	-
Blue	Car	4	Blackwall	-
Green	SUV	4	Whitewall	+
Red	Car	2	Blackwall	+
Green	SUV	2	Blackwall	-
Green	Minivan	4	Whitewall	-

a. Build a decision tree using the above training data using the first 11 instances for training and the last 3 instances for validation. Show the steps for building your decision tree using ID3 algorithm and report the precision and recall on your validation set **(10 points)**

b. Implement a logistic regression classifier (using your own code, not the python packages) and test it on the above data using the same train/validation split. Show the steps for your feature processing and report the precision and recall of your model on the validation set **(15 points)**

c. Implement k-fold cross validation in your logistic regression, and re-apply your model on the above dataset, showing the precision and recall for a 5-fold cross validation run. **(10 points)**

d. Compare and contrast the results you got in parts a,b and c. Explain the differences (if any) from your point of view. **(10 points)**

3. In this problem, you will be using a dataset comprised of a set of posts collected from two newsgroups, one for baseball enthusiasts and the second for hockey fans. The goal is to train a classifier that can accurately predict whether a given post is talking about baseball or hockey. using Naive Bayes algorithm. The dataset provided on Blackboard consists of two files for each newsgroup representing the training and test set for each. There is also a file called “vocabulary.txt” that contains the 5822 actual words in the dictionary. Each line in the train/test file represents a single post from the corresponding newsgroup, with each value at a given position **i** indicating the number of times the word at the **i-th** line in the vocabulary file appears in that post.

   a. Implement a naive bayes model that accepts training and test data in the above format, train a model using the training data, and provide the precision, recall and model accuracy over the test data. Your implementation should also include an output of the model itself, displaying the different probabilities it learned from the training data for each word belonging each class. Your code has to be implemented from scratch, not using any of the existing libraries **(15 points)**

   b. Train a naive bayes model using the provided data and report the learned model (probabilities of each word in the vocabulary belonging to each class), the overall precision, recall and accuracy of the trained model performance on the test data **(10 points)**

   c. Generate another version of the dataset using TF-IDF weights rather than the word frequencies that are included in the provided datasets. Train a naive bayes model using your implementation on the TF-IDF vectors for all the posts to try to predict the newsgroup the post belongs to. Compare the performance of the model using TF-IDF against the model results you got in part b. **(10 points)**