Course: *CMPS 320- Machine Learning* – Professor: *Dr. Osei Tweneboah*

Homework #4 Due on 11/05/2021

Instructions: While discussion with classmates are allowed and encouraged, please try to work on the homework independently and direct your questions to me.

Part A

1. If your model performs great on the training data but generalizes poorly to new instances, what is happening? Discuss two possible solutions.

2. Suppose we want to compute 5-Fold Cross-Validation error on 200 training examples. We need to compute error N_1 times, and the Cross-Validation error is the average of the errors. To compute each error, we need to build a model with data of size N_2, and test the model on the data of size N_3. What are the appropriate numbers for N_1, N_2, N_3?

3. A drug company has developed a classifier for detecting whether a vaccine developed is effective or not.

   We evaluate the classifier on a test set. Here is the confusion matrix. (In the table, E means effective and NE not Effective.)

		Predicted	
		E	NE
Truth	E	970	25
	NE	10	15

   (a) Compute the accuracy of the classifier.

   (b) What is the accuracy of a majority-class baseline? (The class effective (E) is the most common in the training set.)

   (c) Would you say that the classifier is more useful than the majority-class baseline? Explain why or why not.

   (d) Compute the precision of the classifier.

   (e) Compute the recall of the classifier.

   (f) Compute the F_1 score.

Part B

In this homework, the data set used contains information on customers of an insurance company. The data includes product usage data and socio-demographic data derived from zip area codes. The response Variable (Purchase) indicates whether the customer purchased a caravan insurance policy. Each observation corresponds to a postal code. Variables beginning with M refer to demographic statistics of the postal code, while variables beginning with P and A (as well as CARAVAN, the target variable) refer to product ownership and insurance statistics in the postal code. Further information on the individual variables can be obtained at:

http://www.liacs.nl/ putten/library/cc2000/data.html

1. Understanding the Dataset: Given the provided datasets (as CSV files), load them and answer the following questions.

   (a) What is the dimension of the datasets?

   (b) How many predictors measure demographic characteristics?

   (c) What is the percentage of people who purchased caravan insurance?.

2. Data preprocessing: Standardize the data matrix X so that all variables are given a mean of zero and a standard deviation of one. In standardizing the datasets, exclude the response variable.

3. Split the datasets into a test set, containing the first 1,000 observations, and a training set, containing the remaining observations.

   (a) How many observations are in each set?

   (b) How many customers purchased insurance in each set?

4. Binary Classifier: KNN and SGD classifiers

   (a) Apply the K-Nearest Neighbors (KNN) classifier to the caravan dataset. Choose the values $K = 1, 3$ and 5 and for each K, compute the precision and recall. Please comment on the precision.
   Hint: Details of the KNN classifier can be found here:
   https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

   (b) Next apply the Stochastic Gradient Descent (SGD) classifier on the caravan dataset. Compute the precision and recall. Set random seed to 42.

   (c) Which classifier finds real patterns in the caravan dataset?