Assignment 2 Naïve Bayes (75 points)

CSCE 415: Machine Learning

Spring 2021

100 Points

Due: 19 February 2021 11:30 PM

Submission Instructions: Turn in this assignment in a zip or tar file through Blackboard. The zip or tar file should be of the form *last_name_only.zip* Each problem should be in a separate folder before zipping/tarring the two folders.

Problem 1 (50 points): Naïve Bayes Classifier for SPAM

You have been provided the SMS Spam Collection which contains spam messages (labeled as 'spam') and non-spam messages (labeled as 'ham'). Your task is to develop a classifier that best predicts whether a message is spam or not spam; i.e. ham. You will need to prepare the data so that can create a set of words for each record by eliminating punctuation, convert text into lower-case, and split up the words from the text string. Split your data set into a training and testing set. Analyze the results of using Naïve Bayes classification. Follow the process outlined in class. Turn in your Jupyter Notebook (JN) file – do not turn in your data. Make sure you include data visualizations. I will run your JN file, so be sure to only alter your data through Jupyter Notebook.

Problem 2 (50 points): Logistics Regression

For this problem you have file containing 768 records from National Institute of Diabetes and Digestive and Kidney Diseases. This is a subset from a larger database. Your task is to use Logistics Regression to predict an 'Outcome' of diabetes. Some of the records are missing data. Part of this task is to clean up the data. You must use use your Jupyter Notebook with python3, to change any of the records, so that the process can be run against my dataset. Develop the best Logistics Regression classifier that you can after cleaning up the data and analyze your results. Follow the process outlined in class. Explain your choices! Turn in your Jupiter notebook file and your write up. Do Not Turn In The Data! If you change the data in anyway, it must be through your Jupyter Notebook file. Don't forget to include your visualizations.

Please turn in your files showing your work in addition to your write up in Blackboard. Again, put the two problems in different folders, zip or tar them.