STA 4364 HW 3

due Friday October 30 by 11:59PM on Webcourses

Important submission note: All problems must be submitted in either an RMarkdown file or a Jupyter Notebook file (or both, you can submit at most one of each). Any other format will result in a 4-point deduction from the final score.

Problem 1: (10 points) This problem will examine the data set `online_shoppers_intention.csv`. Each row gives a variety of information about a single individual's online habits. Information about the features in this dataset can be found here. We are interested in the `Revenue` response, which is a binary variable indicating whether an individual made a purchase while browsing. Analyze this dataset by following the steps below:

(a) Load the data. Make sure categorical variables are formatted correctly. Merge rare categories for categorical features (40 or fewer observation) into a single feature. Split the data into a train and test set.

(b) Use the training data to create 4 different models to predict `Revenue` given all other features: a logistic regression model, an LDA model, a QDA model, and a Naive Bayes model. Your models can use all features as predictors, for this problem you don't have to investigate feature selection.

(c) Using the validation data, make an ROC curve for each model and calculate the AUC for the ROC curves. Comment on the differences that you observe between the different methods.

(d) Using the validation data, make a Precision-Recall curve for each model, and calculate the AUC for the Precision-Recall curves. Do you notice any different between the results for the ROC curves and the PR curves? What aspect of the response variable `Revenue` likely accounts for this difference? Is the ROC curve or the Precision-Recall curve more reliable for this dataset? Which model(s) do you prefer for prediction in this situation?

Problem 2: (10 points) In this problem, you will build a model to predict the author of a document based on the document text. The dataset is in the file `C50.zip`. This file contains articles written by 50 different authors for Reuters news. There is a training set with 2500 documents in the `C50train` folder (50 articles per author) and a testing set with 2500 documents in the `C50test` (50 articles per author). You will extract word features from the training set (called a "Bag of Words" classification approach) and use this to predict authorship in the test data.

(a) Process the raw text data using the following steps:

- Create a list of keywords from all of the documents in the training data. Remove all numbers and punctuation, change all letters to lowercase, use only a single vocabulary string to represent words with a common stem, and remove common words (known as "stop words") that are contained in the file `stop_words.csv` from your vocabulary list. Sample code for doing this will be provided in class.
- Create two data matrices (one for the training data and one for the validation data) where each rows corresponds to a single document, one column corresponds to author identity and all other columns are binary categorical features indicating if the keyword corresponding to that column is in the text. Remove columns from both the training and validation data matrices that correspond to key words that occur in less than 1% of documents in the training set.

(b) Learn a Naive Bayes classifier for authorship with your training set and predict authorship for documents in the validation set. Report the overall accuracy as well as a confusion matrix for your predictions with each model.

(c) *Bonus (5 points)* Use logistic regression with either L1 or L2 regularization to classify the documents. To do this, you can use a one-vs.-all classfication approach: learn a binary logistic regression classifier for each author (50 classifiers in all), where each classifier is trained on a binary response indicating whether a given document is written by the corresponding author (response is 1) or not (response is 0). When predicting, it is possible that you will have multiple classifiers that output 1 (indicating multiple potential authors), so to make your final prediction choose the author model with the highest probability. It is also possible that no classifier will output a response of 1, in which case you should again choose the author with highest probability as your prediction. Train the models using the training set and report the overall accuracy and confusion matrix for the validation set.