IEOR 4574: Business Analytics for OR
Spring 2016

Professor Adam Elmachtoub
Professor Jacob Leshno

Assignment 3: Classification Methods

Due date: March 3, 11:59pm

Attention: Please prepare two files for each homework assignment: a .pdf file for your answers including relevant figures, and a .R file for your relevant R scripts. File names should be `Last_First_hw.pdf` and `Last_First_hw.R`, e.g., `Obama_Barack_2.pdf` and `Obama_Barack_2.R`. Your submissions must be based on your own original work. Late submissions will not be accepted.

1. In this questions you are asked to analyze consumer data to target a promotion for the MM brand. We will use the data from `OrangeJuice.csv`, which includes observations on customer orange juice purchases. The first variable `Purchase` is the brand of orange juice the consumer previously purchased, which is either the brand MM or CH. The other variables are as following:

   - WeekofPurchase - Week of purchase
   - StoreID - Store ID
   - PriceCH/PriceMM - Price charged for CH/MM
   - DiscCH/DiscMM - Discount offered for CH/MM
   - SpecialCH/ SpecialMM - Indicator of special on CH/MM
   - LoyalCH - A proxy for customer brand loyalty for CH
   - SalePriceCH/SalePriceMM - Sale price for CH/MM
   - PriceDiff - Sale price of MM less sale price of CH

   (a) Load the data from `OrangeJuice.csv` and split your sample into training (50%), validation (25%), and test (25%) data. We will not do model assessment, so no test data. Use the command set.seed(4754) to set the randomizer's seed. Print the summary of the training data.

   (b) Fit a logistic regression to predict `Purchase` using all the covariates over the training data. Print the estimated coefficients and interpret them.
   Hint: Some of the coefficients cannot be estimated, to explain why you should notice that several covariates are can be computed from other covariates.

   (c) Fit a logistic regression that uses only two covariates to predict `Purchase`. Explain why you chose these covariates, and print and interpret the estimated coefficients.

(d) Fit a decision tree to the data to predict `Purchase` using the training data. Plot the tree.

(e) Consider linear support vector classifiers with cost parameters among {0.01,0.1,1,10,100}. What is the best choice for cost parameter based on the validation data? Use the number of correct predictions as your performance measure.

(f) Choose the best model from the 3 previous parts based on the validation data. Which method performed the best?

(g) Evaluate the performance (correct prediction rate) of your best model on the test data.

(h) You are running a promotion aimed to convince customers to sample the MM orange juice. Your campaign wishes to target customers who bought CH orange juice and give them a coupon for MM. Your marketing team tells you that a coupon handed to a customer who bought CH will help convert the customer and will generate a $3 profit, however a coupon that is handed to a customer who already bought MM will be a waste and will generate a loss of $1. Using one of the previous models, you need to decide who receives coupons. For methods that output a probability, find that optimal threshold to convert the probability to a decision. Print the best attainable payoff on the test data.

2. The data in `classifyMe.csv` includes two covariates and a binary outcome.

(a) Notice that the last row of the data is an outlier. Load the data and create two data sets, the first with all rows and the second without the outlier. Print a summary of both.

(b) Fit an LDA classifier to both data sets. Print the LDA models, and describe how they differ.

(c) Fit a linear Support Vector Classifier to both data sets. Describe the differences between the two models.

(d) Use the additional data in `classifyMeValidation.csv` to see the performance of all four estimated classifiers out of sample. Measure performance by number of correct predictions. Rank the performance of the four.

(e) Removing the outlier had a different effect on the two methods, explain why.

3. The file in `cuisinePreferences.csv` contains the cuisine preferences collected in the survey. The first 18 columns are the ratings of the different cuisines, and the last column indicates which section the student belongs to.

(a) Take the data corresponding to your section and use it as your training data. The data from the other section will serve as test data.

(b) Let us say that two students have similar cuisine preferences if their rankings agree on cuisines they both ranked. Use the Euclidian distance on common rankings to create a matrix of the similarity between each two students in your section. Print the 5 closest students to you.

(c) Use the 3-NN method to complete the missing rankings in the data. Create a prediction matrix of how each student in your section will rank each cuisine.

(d) Find the number of neighbors that minimizes the in training RMSE. Consider number of neighbors from 1 to 20 and plot the RMSE (Root MSE) for each.

(e) Using the best number of neighbors you found, run K-NN to predict the cuisine choices for 3 students from the other section. What was the RMSE of the predictions for the 3 students among all cuisines?