IEOR 4650: Business Analytics
Fall 2017 Professor Jacob Leshno

Assignment 3: Classification Methods

Due date: Monday Oct 23, 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., `Skywalker_Luke_2.pdf` and `Skywalker_Luke_2.R`. Your submissions must be based on your own original work. Late submissions will not be accepted.

1. DogBark inc. developed a new dog toy. Because it is a small firm, big retailers would not stock the dog toy until it has been proven to be successful. Therefore, DogBark inc. decided to start by selling the toy directly to consumers. To find interested consumers, the company acquired a dataset of potential customers. However, the dataset includes customers who do not have a dog, and would not be interested in purchasing the dog toy. To try to better target interested the company developed several indexes that may help predict whether the potential customer is a dog owner. To test these indexes, the company surveyed a small subset of customers, asking them whether they have a dog.

   The file `dog.csv` contains the survey data. It contains the following:

   - dog - 'Yes' for dog owners, 'No' for non-dog owners (from survey)
   - pub_dist - the distance from the place of residence to the nearest pub or bar
   - supermaket_dist - the distance from the place of residence to the nearest supermarket
   - laundry_dist - the distance from the place of residence to the nearest laundromat
   - park_dist - the distance from the place of residence to the nearest park
   - neigh_density_score - a score on a scale of 1 to 10 indicating whether the customer lives in a densely populated area
   - tree_score - the number of trees within a circle around the place of residence

   DogBark inc. wants to use this data to mail a pamphlet about the dog toy to customers. Sending each pamphlet costs \$1. Based on previous marketing experience, the company assumes each dog owner who gets the pamphlet will have a 5% chance of purchasing the toy. The toy is made to order, sells for \$50 and costs \$20 to make. Non dog owners who receive the pamphlet will not generate any profit. Unless stated otherwise, answer the questions assuming that DogBark inc. wishes to maximize profits (revenue minus total costs).

(a) Should DogBark send pamphlets to all potential customers?

(b) Load the data from `dog.csv` and split your sample into training (75%) and validation (25%). We will not use a test dataset for this exercise. Use the command set.seed(4650) to set the randomizer's seed. Print the summary of the training data.

One idea is to try to use tree_score to classify potential customers, as it is possible that dog owners would prefer to have many trees close to their place of residence. That is, we will set a threshold and send a pamphlet only to customers whose tree_score is above/below the threshold.

(c) Find the optimal threshold (that maximize profits) for the training data. What is the confusion matrix for the training data? What would have been be DogBark inc.'s profit if it were to use this method to target potential customers within the training data?

(d) Evaluate the classifier from the previous question on the validation data. Is the performance better or worse? Explain why.

(e) How much profits would DogBark inc. make if it had a perfect classifier?

(f) Fit a logistic regression to the training data using all the covariates to predict who is a dog owner. Print the estimated coefficients and interpret them.

(g) Use use the output of the logistic regression to create a classifier. What is the threshold that maximizes DogBark inc.'s profits? What is the confusion matrix?

(h) Fit a decision tree to the training data to predict who is a dog owner. Plot the tree. What is the confusion matrix? What would have been DogBark inc.'s profit if it were to use this method to target potential customers within the training data?

(i) Evaluate the performance of each of the classifiers on the validation data. Which method performed the best in terms of total error rate? Which method would generate the highest profits ? Which method would you recommend using?

(j) Suppose that DogBark inc. got stuck with a large inventory of dog toys, and wishes to change the goal of the market campaign. Instead of maximizing profits, DogBark inc. wishes to use the marketing campaign to get 1,000 purchases by sending the minimal number of pamphlets. DogBark inc. has mailing information and the indexes included in the dataset about 1,000,000 potential customers. The data in `dog.csv` is a representative sample of the larger dataset. Using the classifier you selected in the previous question, how many pamphlets (on average) would DogBark inc. would have to send in order to get 1,000 purchases?

2. 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?