Homework 2

Due: 11.59pm on Friday, January 31

Submission instructions: Submit one write-up per group on gradescope.com.

1 Question

In this question, you will explore prices of used cars as a function of different input variables. You can download data from:
https://github.com/ChicagoBoothML/MLClassData/raw/master/UsedCars/UsedCars.csv

In R, simply run

```
download.file(
   "https://github.com/ChicagoBoothML/MLClassData/raw/master/UsedCars/UsedCars.csv",
   "UsedCars.csv")
```

which will download the file into your current working directory.

1. Take a look at the data-set and describe for what kind of business related problems you could use this data. That is, why would anyone care to collect this data?
2. Split data into two parts: training set consisting of 75% of observations and a test set consisting of 25% of observations.
3. Using ordinary linear regression, find a relationship between `price` and `mileage` of the form

   $$\texttt{price} = b_0 + b_1 \times \texttt{mileage} + e$$

   using the training data. Create a scatter plot of `price` vs `mileage`. Include the best linear regression fit onto the plot.
4. You might notice that the linear fit does not capture the true relationship well. Perhaps we can do better with a polynomial function. Recall polynomial regression from applied regression
   or business statistics class (see also Section 7.1 of Introduction to Statistical Learning) and fit polynomials of the form

   $$\texttt{price} = b_0 + b_1 \times \texttt{mileage} + b_2 \times \texttt{mileage}^2 \ldots + b_d \times \texttt{mileage}^d + e.$$

   Use cross-validation to pick the optimal degree of the polynomial. Plot the cross-validation estimate of the mean-squared error as a function of the degree of the polynomial. Also, report the optimal polynomial degree and add the best fitted polynomial to the scatter plot you created before. Remember to refit the polynomial on all of the training data once you choose the optimal degree.
5. Use k-NN and regression trees to find the relationship between `price` and `mileage`. Again, use cross-validation to find the optimal tuning parameters for these two procedures: k for k-NN and the number of leaves for decision trees. Create plots showing the cross-validation estimate of the mean-squared error as a function of tuning parameters. Create a scatter plot of `price` vs `mileage` that also shows the best polynomial regression, k-NN, and regression tree fit. Which fit would you use and why? Report the test error for the model you select.
6. Now try using `mileage` and `year` to predict `price` using k-NN and regression trees. Remember to rescale data when using k-NN. As before, use cross-validation to pick optimal tuning parameters and plot the cross-validation error curves. How does the optimal k change as compared to when you were using only `mileage`? How about the size of the optimal tree? Does your model perform better when including `mileage` as an additional variable?

7. Finally, run a regression tree using all the variables to predict `price`. Use cross-validation to select the optimal size of the regression tree. Report the mean-squared error on the test set for the optimal tree.

OPTIONAL BONUS QUESTION: We can use cross-validation to select relevant variables for predicting the price of a used car. Try finding out whether all variables are predictive using regression trees and cross-validation. Think about and describe how would you try to find a simpler model, that is, one that does not include all the variables.

2 Question

Throughout this question you will use files `Wine_deals.csv` and `Wine_transactions.csv`.

A business imports wine in large quantities and sells is to wine and liquor stores. Each month the business creates two or three deals on different wines. For example, in January there were two deals, one on Malbec and one on Pinot Noir. Last year the business offered 32 deals, which are described in the file `Wine_deals.csv`. For each deal, there is information on the type of wine offered, minimum number of bottles needed to qualify for the discount, discount percentage off of the retail price, origin of wine and whether the wine has passed its peak. The business also has information on customers and deals they bought in the file `Wine_transactions.csv`. For example, person with the last name Smith bought some Pinot Noir in January and another deal on Pinot Noir in September.

The business would like to use this data to better understand their customers. They have heard that Chicago Booth students are amazing data analysts and decided to hire your group to extract useful information from data. You have decided to first cluster data and look for similar customers.

Represent each customer using the previous deals they purchased. Run k-means clustering for several values of k. The goal is to find a reasonable number of clusters you could act on and that you could explain reasonably well. Tell a story, describe a typical customer in each of the clusters. For example, customers may like certain type of wine, certain origin, large/small quantities, large discounts, etc. Sometimes it is not easy to describe clusters. This is actually often going to be the case, as it is rarely the case that we can perfectly separate customers. There are two ideas you could use in order to better understand customers. First, look at the centers of each cluster. Second, identify offers that members of each cluster bought. This will help in understanding clusters.