Homework 1

Due April 12, 2019 by 11:59pm

Instructions: **All code exercises must be completed using Python.** Upload your answers to the questions below to Canvas. Submit the answers to the questions (including the relevant output of the code) in a PDF file and your code in a (single) separate file. Be sure to comment your code to indicate which lines of your code correspond to which question part.

1. Read Chapter 2 in *Elements of Statistical Learning*.

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

   (a) You want to predict whether a particular customer is going to click on an online advertisement or not. You have information on whether or not they clicked on 200 other ads, in addition to whether the ad was in the same category, whether the ad was shown during regular working hours, whether the ad was shown on a weekend, and the percent of all customers who had previously clicked on the ad.

   (b) Suppose it is the end of the quarter and you wish to predict your score on the final exam. You have data from 20 classes you have previously taken, consisting of your final exam scores, your average scores on the midterms (i.e., one average midterm score per class), your average homework scores (i.e., one average homework score per class), and whether the final exam was take-home or not.

   (c) You work for an ice cream shop and are in charge of determining what factors affect how much ice cream is sold each day. For 300 days you have information on how much ice cream the shop sold, in addition to whether the day was sunny or not, what the temperature was, whether school is in session or not, whether your most popular flavor was available that day, and whether you had recently run any advertisements.

3. In this problem you will brainstorm real-life applications for statistical learning. Your answers aren't allowed to be the same as any of the examples in the other homework problems.

   (a) Describe three real-life applications in which classification might be useful, **one from political science, one from sports, and one from an area of your choice**. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(b) Describe three real-life applications in which regression might be useful, **one from engineering, one from business, and one from an area of your choice**. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

(c) Describe three real-life applications in which cluster analysis might be useful, **one from education, one from meteorology, and one from an area of your choice**. Be sure to describe why it would be useful.

4. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Suppose we wish to use this data set to

Obs.	X_1	X_2	X_3	Y
1	0	4	0	Green
2	2	0	1	Red
3	0	1	3	Red
4	-1	1	2	Green
5	-3	0	1	Green
6	2	0	1	Red

make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is linear but the data is noisy, then would we expect the best value for K to be larger or smaller? Why?

5. This exercise relates to the `College` data set, which can be found here `http://www-bcf.usc.edu/~gareth/ISL/College.csv`. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- `Private`: Public/private indicator
- `Apps`: Number of applications received
- `Accept`: Number of applicants accepted
- `Enroll`: Number of new students enrolled
- `Top10perc`: New students from top 10% of high school class
- `Top25perc`: New students from top 25% of high school class
- `F.Undergrad`: Number of full-time undergraduates
- `P.Undergrad`: Number of part-time undergraduates
- `Outstate`: Out-of-state tuition

- `Room.Board`: Room and board costs
- `Books`: Estimated book costs
- `Personal`: Estimated personal spending
- `PhD`: Percent of faculty with Ph.D.s
- `Terminal`: Percent of faculty with terminal degree
- `S.F.Ratio`: Student/faculty ratio
- `perc.alumni`: Percent of alumni who donate
- `Expend`: Instructional expenditure per student
- `Grad.Rate`: Graduation rate

Before reading the data into Python, it can be viewed in Excel or a text editor.

(a) Use the `pandas.read_csv()` function to read the data into Python. Call the loaded data `college`. Make sure that you have the directory set to the correct location for the data if the file is saved on your computer.

(b) Look at the data using the `head` attribute. You should notice that the first column is just the name of each university. We don't really want Python to treat this as data. However, it may be handy to have these names for later. Try the following commands:

```
college.rename(columns={'Unnamed: 0': 'School'}, inplace=True)
college.set_index('School')
```

(The line before 0 denotes a space.) You should see that there is now a `School` column with the name of each university recorded. This means that Python has given each row a name corresponding to the appropriate university. Python will not try to perform calculations on the row names.

(c) i. Use the `describe` attribute to produce a numerical summary of the variables in the data set.

ii. Use the `scatter_matrix()` function from the package `pandas.plotting` to produce a scatterplot matrix of the second through fourth columns of the data.

iii. Use the `boxplot` attribute to produce side-by-side boxplots of `Room.Board` versus `Private`. Hint: Use the `column` option to select the continuous variable and the `by` option to select the categorical variable.

iv. Create a new qualitative variable, called `Elite`, by binning the `Top10perc` variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

```
Elite = np.array([False]*len(college))
Elite[college['Top10perc'] > 50] = True
college['Elite'] = pd.Series(Elite, index=college.index)
```

Use the `describe` attribute to see how many elite universities there are. For this you might find the option `include=['bool']` useful. Now use the `boxplot` attribute to produce side-by-side boxplots of `Room.Board` versus `Elite`.

v. Use the `hist` attribute to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the `layout` option useful: it will divide the figure into regions so that plots can be made simultaneously.

vi. Continue exploring the data, and provide a brief summary of what you discover.

6. This exercise involves the `Auto` data set found here: `http://www-bcf.usc.edu/~gareth/ISL/Auto.data`. Make sure that the missing values have been removed from the data.

   (a) Which of the predictors are quantitative, and which are qualitative?

   (b) What is the range of each quantitative predictor?

   (c) What is the mean and standard deviation of each quantitative predictor?

   (d) Now remove the last 35 observations. What is the range, mean, and standard deviation of each quantitative predictor in the subset of the data that remains?

   (e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

   (f) Suppose that we wish to predict gas mileage (`mpg`) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting `mpg`? Justify your answer.