Assignment 1

CS 750/850 Machine Learning

- **Due**: Monday 2/3 at 11:59PM
- **Submisssion**: Turn in as a **PDF** and the **source code** (R,Rmd,py,ipynb) on MyCourses
- **Questions**: Piazza and Office hours: *Marek*: Wed 1:30-3:00pm, Soheil: Mon 2-4pm, Xihong: Thu 1:30-3:30pm
- **Extra credit**: Especially good questions or helpful answers on Piazza regarding the assignment earn up to 5 points extra credit towards the assignment grade.

The instructions are geared towards R users. The comments in [P: xxx] are meant as hints to Python users.

Feel free to achieve the results using commands other than the ones mentioned. R, especially, shines when it comes to processing and visualization of structured data, but you would not be able to tell from the ISL book. It uses the oldest and simplest (and ugliest) subset of R. I recommend checking out `dplyr` for data processing [3,4] and GGPlot [1,4] for plotting.

Problem 1 [25%]

In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use `set.seed(1)` [P: `np.random.seed(1)`] prior to starting part (1) to ensure consistent results.

1. Using the `rnorm()` [P: `np.random.normal`] function, create a vector, `x`, containing 100 observations drawn from a $\mathcal{N}(0,3)$ distribution (Normal distribution with the mean 0 and the **standard deviation** $\sqrt{3}$). This represents a feature, X.
2. Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $\mathcal{N}(0,0.5)$ distribution i.e. a normal distribution with mean zero and standard deviation $\sqrt{0.5}$.
3. Using `x` and `eps`, generate a vector `y` according to the model Y:

   $$Y = -2 + 0.6X + \epsilon$$

   What is the length (number of elements) of `y`? What are the values of β_0, β_1 in the equation above (intercept and slope)?
4. Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe. [P: see [2]]
5. Fit a least squares linear model to predict `y` using `x`. Comment on the model obtained. How do $\hat{\beta}_0, \hat{\beta}_1$ compare to β_0, β_1?
6. Display the least squares line on the scatterplot obtained in 4.
7. Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

Optional Problem O1 [30%]

This problem can be substituted for Problem 1 above, for up to 5 points extra credit. At most one of the problems 1 and O1 will be considered.

Read Chapter 1 and solve Exercises *1.6* and *1.10* in [Bishop, C. M. (2006). Pattern Recognition and Machine Learning].

Problem 2 [25%]

Read through Section 2.3 in ISL. Load the `Auto` data set and *make sure to remove missing values from the data*. Then answer the following questions:

1. Which predictors are *quantitative* and which ones are *qualitative*?

2. What is the range, mean, and standard deviation of each predictor? Use `range()` [`pandas.DataFrame.min` and `max`] function.
3. Investigate the predictors graphically using plots. Create plots highlighting relationships between predictors. See [1] for a ggplot cheatsheet.
4. Compute the matrix of correlations between variables using the function `cor()` [P: `pandas.DataFrame.corr`]. Exclude the `name` variable.
5. Use the `lm()` function to perform a multiple linear regression with `mpg` as the response. [P: using `rpy` package is acceptable] Exclude `name` as a predictor, since it is qualitative. Briefly comment on the output: What is the relationship between the predictors? What does the coefficient for `year` variable suggest?
6. Use the symbols `*` and `:` to fit linear regression models with interaction effects. What do you observe?
7. Try a few different transformations of variables, such as $\log(X)$, $\sqrt{X}$, X^2. What do you observe?

Problem 3 [25%]

Using equation (3.4) in ISL, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.

Problem 4 [25%]

It is claimed in the ISL book that in the case of simple linear regression of Y onto X, the R^2 statistic (3.17) is equal to the square of the correlation between X and Y (3.18). Prove that this is the case. For simplicity, you may assume that $\bar{x} = \bar{y} = 0$.

References

Each reference is a link. Please open the PDF in a viewer if it is not working on the website.

1. R GGPlot cheat sheet
2. Python Pandas data visualization
3. R For Data Science
4. Cheatsheets