Assignment 1

CS780/880: Introduction to Machine Learning

Due: By 12:40PM Tue Feb 14th, 2017
Submission: Turn in as a PDF on myCourses, or printed and turned in at class; if other methods fail, email to mailto:mpetrik@cs.unh.edu with Subject that contains the string `[CS780880HW]`
Discussion forum: https://piazza.com/unh/spring2017/cs780cs880
Applied problems: Install and learn to use R (https://www.r-project.org/), read the labs in ISL. It is recommended to use R Notebooks of RStudio to solve and typeset homeworks.

Problem 1 [10%] What are the advantages and disadvantages of very flexible (vs less flexible) approach for regression or classification? When would be a more flexible approach preferable? What about a less-flexible approach?

Problem 2 [10%] Describe some real-life applications for machine learning.

1. Describe one real-life application in which *classification* combined with *prediction* may be useful. Describe the response and predictors.
2. Describe one real-life application in which *classification* combined with *inference* may be useful. Describe the response and predictors.
3. Describe one real-life application in which *regression* combined with *prediction* may be useful. Describe the response and predictors.
4. Describe one real-life application in which *regression* combined with *inference* may be useful. Describe the response and predictors.

Problem 3 [35%] 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)` prior to starting part (a) to ensure consistent results.

1. Using the `rnorm()` function, create a vector, `x`, containing 100 observations drawn from a $N(0,1)$ distribution. This represents a feature, X.

2. Using the `rnorm()` function, create a vector, `eps`, containing 100 observations drawn from a $N(0,0.25)$ distribution i.e. a normal distribution with mean zero and variance 0.25.

3. Using `x` and `eps`, generate a vector `y` according to the model Y:

$$Y = -2 + 0.75X + \epsilon$$

   What is the length of `y`? What are the values of β_0, β_1?

4. Create a scatterplot displaying the relationship between `x` and `y`. Comment on what you observe.

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.

Problem 4 [35%] 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 (and show your code):

1. Which predictors are *quantitative* and which ones are *qualitative*?
2. What is the range, mean, and standard deviation of each predictor? Use `range()` function.
3. Investigate the predictors graphically using plots. Create plots highlighting relationships between predictors.
4. Compute the matrix of correlations between variables using the function `cor()`. Exclude the `name` variable.
5. Use the `lm()` function to perform a multiple linear regression with `mpg` as the response. Exclude `name` as a predictor, since it is qualitative. 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?

CS880 Graduate: Problem 5 [10%] 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 $\bar{x} = \bar{y} = 0$.

CS780 Undergraduate: Problem 5 [10%] Using equation (3.4) in ISL, argue that in the case of simple linear regression, the least squares line always passes through the point $(\bar{x}, \bar{y})$.