DS 5220 / CS6140-02 - Fall 2017 Olga Vitek

Homework 2

Each part of the problems 5 points
Due on Blackboard before midnight on Tuesday October 3, 2017.

1. List all members of your project group: first and last name, program of study, and year in the program.

2. In this question, we will revisit JWHT Figure 3.3, in terms of estimation of parameters of linear regression. Set random seed to 123.

   (a) Simulate N=100 values of X_i, distributed Uniformly on interval (-2,2). Simulate the values of $Y_i = 2 + 3X_i + \varepsilon_i$, where ε_i is drawn from $\mathcal{N}(0, 4)$. Estimate the slope of linear regression using least squares. How close is it to the truth?

   (b) Implement batch gradient descent and stochastic gradient descent to estimate the slope and the intercept, as function of the learning rate α. Stop the iterations when the change in the parameter between two consecutive iterations is less than 0.1, or when the number of iterations exceeds a large number (say, 1000).

   (c) Consider a range of parameter α. Estimate the parameters of the regression using batch gradient descent, and using stochastic gradient descent. Plot the estimates of the slope as function of α. Plot the time until convergence as function of α. Interpret the result, and suggest the best α. [Note: try a few ranges of α to get meaningful results]

   (d) Repeat (a) 200 times. Each time, record the slope of the regression estimated by least squares, by batch gradient descent with best α from (c), and by stochastic gradient descent from best α from (c). Plot three histograms of these estimates, and overlay the true value. Plot the histograms of time until convergence. Comment on the results.

   (e) Repeat (d), while changing the number of observations N to 300.

3. In this question we will revisit Simpsons paradox, using the `diamonds` dataset in the R package `ggplot2`. Access the dataset in R by typing `library(ggplot2); data(diamonds)`. We will only consider the variables `price`, `color` and `carat`. [Note: "Color" is a continuum, coded by letters in alphabetic order. "Carat" is a measure of diamond's size and weight.]

   (a) Plot `price` vs `color`, and comment on the relationship. [Hint: use boxplots.]

   (b) Partition diamonds into 5 groups according to the value of `carat`, such that there is a same number of observations in each group. (I.e., the first group contains diamonds with `carat` between the minimal value and 20th quantile, the second

group between 20th and 40th quantile etc). Plot `price` vs `color` for separate group, and explain the reasons for the relationship change.

(c) Fit a linear regression that predicts `price` as function of `color` and `carat` and justify your choice. Visualize the model fit. [Hint: in addition to Simpsons paradox, consider other issues such as non-linearity and unequal variance]

4. **(35pts)** The dataset `credit` was used by JWHT to illustrate the use of linear regression, e.g. in Chapter 3.3.1. Download this dataset from `http://www-bcf.usc.edu/~gareth/ISL/data.html`. See JWHT Ch 3.3.1 for details.

   In this question, we will perform the full linear regression analysis of this dataset, to predict the value of `balance` on the credit card. Set random seed to 123 and perform the following steps:

   (a) **Select the training set:** Randomly select 200 subjects into the training subset of the data.

   (b) **Data exploration:** Consider the training set only. Report one-variable summary statistics, two-variable summary statistics (e.g., correlations). Discuss the implications of the exploration for the regression analysis (e.g., presence of highly correlated predictors, categorical predictors, missing values, outliers etc).

   (c) **Assumption of Normality:** Consider the training set only. Fit linear regression with all predictors. Evaluate the plausibility of Normal linear regression and constant variance using a quantile-quantile plot of the residuals obtained from the model with all the possible predictors. Transform `balance` if needed. [Hint: for simplicity, here consider the additive model only]

   (d) **Variable selection:** Consider the training set only. Perform variable selection using all subsets selection. [Hint: it may be interesting to also consider statistical interactions]

   (e) **Variable selection:** Consider the training set only. Perform variable selection using statistical regularization. [Hint: it may be interesting to also consider statistical interactions]

   (f) **Performance evaluation:** Evaluate the performance of the models selected in the items above, using the predictive accuracy on the validation set. Which model performs best, and why?

   (g) **Interpretation of the results:** Interpret the model with the best fit, using both English language description, and data/model visualization of your choice.