CS 76140 - Spring 2017 Olga Vitek

Homework 2

Due on Blackboard before 10am on Thursday February 9, 2017.

Note: Use any tool of your choice (including Word, latex, Markdown, or pencil+paper) to prepare the solutions. The answers should be easy to find and grade. Unreadable hand-written solutions will be given 0 points, at the grader's discretion, regardless of the correctness of the answer. For each problem, use the appropriate notation for random variables, probabilities etc. State the full formula in addition to the numerical conclusions. For data analysis problems, use reproducible research tools such as R Markdown whenever possible.

1. **(5pts)** Looking ahead to the end-of semester project, we need to form groups of your project collaborators. Each group should consist of 3-4 students. Smaller groups are allowed, but need to be approved by the instructor in advance of the homework deadline. In response to this question, list your project group collaborators: first and last names, program, year in the program. **Please work individually on all the remaining problems in this homework (i.e., not with your project group).**

2. **(20pts)** Two manuscripts are posted on Piazza:

   (a) S. Ray *et al.*. "Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins". *Nature Medicine*, 13:1359, 2007.

   (b) T. Knickerbocker *et al.* "An integrated approach to prognosis using protein microarrays and nonparametric methods". *Molecular Systems Biology*, 3:123, 2007.

   Choose one the two manuscripts above. Summarize the author's work, by answering the questions below. The summary should not exceed 1 page, 11 points font single-spaced.

   (a) **Description of the problem:** 1-2 paragraphs describing the question addressed by the manuscript, statistical methods used, and major conclusions.

   (b) **Positive aspects:** 1-2 paragraphs describing aspects of the analysis that you think are interesting and well done.

   (c) **Negative aspects:** 1-2 paragraphs describing aspects of the analysis that you think were incomplete or sub-optimal. *[Hint: a major flaw of the Knickerbocker's work is the lack of the validation set.]*

   (d) **Possible extensions:** If given an opportunity and time, how would you improve/extend this work? Would you be able to answer additional questions with this dataset, or use alternative methods to improve the analysis?

3. **(20pts)** The website `http://cs229.stanford.edu/projects2015.html` contains projects of the Machine Learning course at Stanford in Fall 2015. Choose one of the projects. Summarize the author's work, by answering the same questions as in the problem above. The summary should not exceed 1 page, 11 points font single-spaced. **Start the summary with the number and the url of your chosen project.**

   [Note: the Stanford course discussed different topics, and in different depth, than our class. I do not expect you to research the details of the methods that that students used. A general description of the goal and of the positive and negative aspects of the analysis is sufficient for this homework. Post questions on Piazza if you have doubts on the extent of the details.]

4. **(35pts)** The dataset *prostate* was used by Hastie, Tibshirani and Friedman to illustrate the use of linear regression, e.g. in Chapter 3. The dataset is posted on Piazza, and is also available at `http://statweb.stanford.edu/ tibs/ElemStatLearn`. See Chapter 1 of HTF for the description.

   In this question, we will perform the full linear regression analysis of this dataset, to predict the value of `lpsa`. Perform the following steps:

   (a) **Select the training set:** Download the data. Select the training subset of the data, by choosing the rows where the last column ('train') is TRUE.

   (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 using a quantile-quantile plot of the residuals obtained from the model with all the possible predictors. [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.

   (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.