DS 5220 - Fall 2018 Olga Vitek

# Homework 2

Analytical questions are best prepared using Latex/word, however photos of handwritten notes are also acceptable. For the questions involving programming, use a single notebook (Jupyter or R notebook) to answer all the programming questions. Run the code, explain the findings through markdown or visualizations and export it to PDF. Merge the notebook PDF with the rest of the files (latex or photos) of the homework. Submit the single PDF file through Blackboard.

**Due on Blackboard before midnight on Tuesday October 2, 2018.**
Each part of the problems 5 points

1. *[Analytical question]* For a regression in the form $\mathbf{y} = \mathbf{X}\theta + \varepsilon$, derive the ridge regression parameter estimates $\hat{\theta}$ as function of the regularization parameter $\lambda$.

2. *[Analytical question]* Let $\hat{\theta}^{LS}$ be the least squares estimator, and $\hat{\theta}^{Ridge}$ be the ridge regression estimator. Prove that $Var(\hat{\theta}^{LS}) \geq Var(\hat{\theta}^{Ridge})$.

3. *[Analytical question]* For a regression in the form $\mathbf{y} = \mathbf{X}\theta + \varepsilon$, consider a constrained optimization of the form

$$\hat{\theta} = arg\ min_{\theta} ||\mathbf{y} - \mathbf{X}\theta||_2^2, \text{ such that } \mathbf{w}^T\theta = \mathbf{b}$$

for a given $\mathbf{w}$ and $\mathbf{b}$. Assume that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$, where $\mathbf{I}_p$ is the identity matrix. Derive the expression for the estimates $\hat{\theta}$.

4. *[Implementation question]* In this question we will evaluate the ability of regularization to perform variable selection.

   (a) For linear regression with multiple predictors, implement (i) the analytical solution for least squares regression, (ii) the analytical solution for ridge regression as function of the regularization parameter $\lambda$, and (iii) a batch gradient descent optimization for lasso regression as function of the regularization parameter $\lambda$.

   (b) Similarly to homework 1, simulate N=50 values of $X_i$, distributed Uniformly on interval (-2,2). Simulate the values of $Y_i = 2 + 3X_i + e_i$, where $e_i$ is drawn from $\mathcal{N}(0, 4)$. Fit to the simulated data a polynomial regression of degree 5 (i.e., including as predictors $X, X^2, \ldots, X^5$) using the implementations (i), (ii) and (iii) above, with a fixed value of $\lambda$, say $\lambda = 5$.

   (c) Repeat (b) 1,000 times, plot the histograms of the coefficient associated with $X$, and overlay the true value. Plot the histograms of predictions $\hat{Y}$ for $X = 1.5$, and overlay the true expected value.

   (d) Repeat (b) for a range of $\lambda$ (one iteration for each $\lambda$). Plot the coefficient associated with $X$ as function of $\lambda$, and overlay the true value. Plot the predictions $\hat{Y}$ for $X = 1.5$ as function of $\lambda$, and overlay the true expected value.

   (e) Summarize the results, and comment on how the algorithm affects the estimates.

5. *[Case study]* The dataset for this homework is posted on Piazza. It documents residential sales that occurred during the year 2002 in a city in the midwest. Data on 522 arms-length transactions include sales price, style, finished square feet, number of bedrooms, pool, lot size, year built, air conditioning, and whether or not the lot is adjacent to a highway. The city tax assessor is interested in predicting sales price based on the demographic variable information given above.

   The goal of this problem is to develop a model, justify the model choice, and evaluate its use as a tool for predicting sales price. You can use your own implementation of linear regression above, or an existing software package. The problem may have multiple reasonable solutions.

   Perform the following steps:

   (a) **Select the training set:** Randomly select 200 subjects into the training set, 200 for the model development set, and the remainder in the model evaluation set.

   (b) **Data exploration:** Consider the training set only. Report one- and two-variable summary statistics (e.g., boxplots, histograms, scatterplots, correlations - limit the total number of plots, and only present those that are necessary). Discuss the implications of the exploration for the regression analysis (e.g., presence of highly correlated predictors, categorical predictors, missing values, outliers etc). Transform variables if needed.

   (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 plot of the residuals obtained from the model with all the possible predictors. Transform *Sales price* if needed.

   (d) **Variable selection:** Perform variable selection using all subsets selection, by fitting the models on the training set and selecting the best performing subset using the model development set. *[Hint: Consider appropriate coding for qualitative predictors].*

   (e) **Variable selection:** Consider the training set only. Perform variable selection using ridge regularization.

   (f) **Variable selection:** Consider the training set only. Perform variable selection using lasso regularization.

   (g) **Performance evaluation:** Evaluate the performance of the models selected with all subsets, lasso and ridge regularization above, using the predictive accuracy on the validation set. Which model performs best, and why?

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