The University of Texas at Austin
Department of Electrical and Computer Engineering

EE379K: Data Science Lab — Fall 2017

Lab Six

Caramanis/Dimakis Due: Monday Oct 16th, 3:00pm 2017.

Problem 1. In this problem we will use synthetic data sets to explore the bias-variance tradeoff incurred by using regularization.

- Generate data of the form:

$$\boldsymbol{y} = X\beta + \epsilon,$$

where X is an $n \times p$ matrix where $n = 51$, $p = 50$, and each $X_{ij} \sim N(0, 1)$. Also, generate the noise according to $\epsilon_i \sim N(0, 1/4)$. Let β be the all ones vector (for simplicity).

By repeatedly doing this experiment and generating fresh data (fresh X, and y, and hence ϵ) but keeping β fixed, you will estimate many different solutions, $\hat{\beta}$. Estimate the mean and variance of $\hat{\beta}$. Note that $\hat{\beta}$ is a vector, so for this exercise simply estimate the variance of a single component.

- Use ridge regression, i.e. ℓ_2 regularization. Vary the regularization coefficient $\lambda = 0.01, 0.1, 1, 10, 100$ and repeat the above experiment. What do you observe? As you increase λ is the model becoming more simple or more complex? As you increase λ is performance becoming better or worse?

Problem 2. Problem 9 from Chapter 6.
(Predicting the number of applications in College) Note that you will have to read about PCR (Principal Components Regression) and PLS (Partial Least Squares) in the book, since we did not discuss these in class.

Problem 3. Problem 11 from Chapter 6.
(Predicting crime in Boston)
Problem 4. (+20 points extra credit). This is a written problem, supporting Problem 9 above.

- Consider the Least Squares optimization problem, given data X and y:

$$\min_{\beta} : \ \|X\beta - y\|_2^2 = \sum (x_i\beta - y_i)^2.$$

Note that x_i represents the i^{th} row of X and hence is a row-vector. Hence $x_i\beta$ represents the dot product between the p-length vectors x_i and β. Derive a closed form solution (as we did in class) for $\hat{\beta}_{LS}$, by expanding out, taking the derivative and setting it equal to zero. It might be easiest to work in vector notation rather than deal with the individual x_i's.

- Now consider the Ridge Regression problem:

$$\min_{\beta} : \ \|X\beta - y\|_2^2 + \lambda\|\beta\|_2^2 = \sum_i (x_i\beta - y_i)^2 + \lambda \sum_i \beta_i^2.$$

Use the same approach as above to again derive a closed form expression for the solution, $\hat{\beta}_R$.