

Statistical Learning, Spring 2013

Homework exercise 2

Due date: 23 April 2013

Read the note on quantile regression posted on the class website before answering questions 1,2.

1. Playing with quantile regression

Divide the competition data to 8000 training and 2000 validation observations, and fit linear quantile regression for quantiles $\{0.1, 0.25, 0.5, 0.75, 0.9\}$. Apply each model to the validation data, calculate and plot the following, and comment on the results:

- (a) Validation RMSE for each model.
- (b) Average prediction for each model.
- (c) The predictions of all models at five randomly selected validation observations.

Tip: A function to run quantile regression (including absolute loss regression) in R is `rq` in package `quantreg`.

2. Robustness of absolute loss regression and its variants compared to least squares

- (a) **Theoretical:** Assume we have a sample $X = \{x_1, \dots, x_n\}$. Given a statistic $S(X)$, the *breakdown point* is defined as the minimal % of samples that have to be contaminated (in any way we wish) so that $S(X)$ can be *arbitrarily* corrupted (i.e., driven to $\pm\infty$). Show that the breakdown point for the mean $S_1(X) = \bar{x}$ is $1/n$ whereas for the median $S_2(X) = \text{median}(X)$ the breakdown is about $1/2$. What would be the breakdown point for other quantiles?
- (b) **Practical:** Demonstrate the lack of robustness of linear regression compared to absolute loss regression by applying them both to the Prostate cancer dataset and plotting the effect that extreme contamination of one y value in the training set has on the test set RMSE, when each of these modeling approaches is used to calculate $\hat{\beta}$.

3. Ways of interpreting and calculating Ridge and Lasso regression:

- (a) **ESL 3.7:** Show that if we assume a likelihood $y_i \sim N(x_i^T \beta, \sigma^2)$ for $i = 1, \dots, n$ and a prior $\beta \sim N(0, \tau^2 I)$, then the negative log-posterior density of β is $\sum_{i=1}^n (y_i - x_i^T \beta)^2 + \lambda \sum_{j=1}^p \beta_j^2$ up to multiplication and addition of constants, with $\lambda = \sigma^2/\tau^2$. Conclude that the Ridge solution is a maximum posterior estimate of β .
- (b) Show that the same applies to Lasso, except that the prior on β is a double exponential.
 Note: A double exponential random variable with parameter θ has density $f(x) = \theta/2 \cdot \exp(-|x|\theta)$.
- (c) **ESL 3.10:** Show that the ridge regression estimate can be obtained by ordinary least squares regression on an augmented data set. We augment the centered matrix X with p additional rows $\sqrt{\lambda} I_{p \times p}$, and augment y with p zeros. Comment briefly on how we can think of Ridge shrinkage as adding more “neutral” observations with 0 response.
- (d) What would be a corresponding case for the Lasso penalty, where the shrinkage can be accomplished by adding data and solving the same fitting problem?
 Hint: Think beyond squared error loss.

4. **Short intuition questions:**

- (a) If I believe that only a small number of my variables are important, which one (or more) of these four regularization approaches should I use?
- i. Ridge
 - ii. Lasso
 - iii. Variable selection
 - iv. PCA regression
- (b) Same question, except that now I believe that only a low-dimensional linear subspace of the span of my variables is important.