

## Homework exercise 2

Due date: 19 December 2024 before class

### 1. Playing with quantile regression

The note on quantile regression posted on the class website can help in answering this question. 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:

- Validation RMSE for each model.
- Average prediction for each model.
- 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. Ways of interpreting and calculating Ridge and Lasso regression:

- 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  $\beta$  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  $\beta$ .
- Show that the same applies to Lasso, except that the prior on  $\beta$  is a double exponential. Note: A double exponential random variable with parameter  $\theta$  has density  $f(x) = \theta/2 \cdot \exp(-|x|\theta)$ .
- ESL 3.10 (3.12 in 2nd ed.):** 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.
- 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.

### 3. The effect of identical predictors in Ridge and Lasso (3.28, 3.29 in ESL 2nd ed.):

Assume we have a univariate model with one  $x$  variable and no intercept. We fit constrained ridge regression and lasso with a given constraint  $s$  on the norm ( $\ell_2$  norm squared for ridge,  $\ell_1$  norm for lasso). Now we add a second identical variable  $x^* = x$  and refit the models with the same constraint. What happens to the coefficients  $\hat{\beta}$  of both models? How does the two-dimensional solution to the new problem relate to the one-dimensional solution to the old one in each case? Is it unique? Assume the constraint is much smaller than the norm of the least squares solution, so it is tight.

Hint: The behavior of ridge and lasso under this scenario is quite different. Since both predictors  $x, x^*$  are identical, a coefficient can be divided between them in different ways which give the same fit. Consider what different divisions do to the norm of the coefficient vector in each case, and use that to infer the optimal solution. You can also simulate to gain intuition.

#### 4. Guaranteed error reduction via Ridge Regression

Assume the linear model is correct, i.e.,  $E(Y|X = x) = x^T \beta$ . Consider making a prediction at a new point  $x_0$  based on a Ridge Regression with smoothing parameter  $\lambda$ :  $\hat{Y} = x_0^T \hat{\beta}^{\text{ridge}}(\lambda)$

- (a) Derive explicit expressions for the bias and variance of  $\hat{Y}$  as a function of  $\lambda$  (use the SVD of  $X$  for the variance).
- (b) Set  $MSE(\lambda) = \text{bias}^2(\lambda) + \text{Var}(\lambda)$  from above, show that

$$\left. \frac{d}{d\lambda} MSE(\lambda) \right|_{\lambda=0} < 0$$

Suggested approach:

- i. Show by differentiation that  $\left. \frac{d}{d\lambda} \text{Var}(\lambda) \right|_{\lambda=0} < 0$ .
- ii. Show that  $\left. \frac{d}{d\lambda} \text{bias}^2(\lambda) \right|_{\lambda=0} = 0$ . Look at the expression for bias to find a simple argument, avoid complex differentiations!
- (c) Briefly explain the meaning of this result — what happens when we add a *little* ridge penalty to standard linear regression?

Surprisingly, the same is true for the Lasso. The proof, however, is much more involved.

#### 5. 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.