Statistical Learning - Vienna, Fall2017

Homework problems 5+6

5 **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 $i = 1, ..., n$ 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 (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.

(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.

6 **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 λ: $\hat{Y} = x_0^T \hat{\beta}^{\text{ridge}}(\lambda)$

(a) Derive explicit expressions for the bias and variance of $\hat{Y}$ as a function of λ (use the SVD of X for the variance).

(b) Set $MSE(\lambda) = \text{bias}^2(\lambda) + \text{Var}(\lambda)$ from above, show that

$$\frac{d}{d\lambda} MSE(\lambda)\bigg|_{\lambda=0} < 0$$

Suggested approach:

i. Show by differentiation that $\frac{d}{d\lambda}\text{Var}(\lambda)|_{\lambda=0} < 0$.

ii. Show that $\frac{d}{d\lambda}\text{bias}^2(\lambda)|_{\lambda=0} = 0$. 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.