Homework 10

Due Friday, 11/22/19. Please place give your homework to the TA Jiaxuan Duan or place it in his mailbox by Friday afternoon.

1. (Training vs. test error) Suppose $y_i = f(\boldsymbol{x}_i) + \epsilon_i$ for some unknown function f, where $\boldsymbol{x}_i$ is non-random, $\mathbb{E}(\epsilon_i) = 0$ and $\mathbb{E}\left(\epsilon_i^2\right) = \sigma^2$ for $i = 1, \ldots, n$. Suppose you use the training set $\mathcal{T} = \{y_1, \ldots, y_n\}$ to obtain $\hat{f}$, an estimate for f. Define the in-sample **test error** and **training error** to be

$$\text{Err}_{\text{in}} = \mathbb{E}\left[n^{-1}\sum_{i=1}^{n}\left\{\tilde{y}_i - \hat{f}(\boldsymbol{x}_i)\right\}^2 \mid \mathcal{T}\right], \quad \overline{\text{err}} = n^{-1}\sum_{i=1}^{n}\left\{y_i - \hat{f}(\boldsymbol{x}_i)\right\}^2$$

where $\tilde{y}_i$ is an independent copy of y_i (i.e. $\tilde{y}_i$ and y_i are independent and have the same distribution). Err_{in} determines how the estimator performs in the prediction of new unobserved data. Note that Err_{in} and $\overline{\text{err}}$ are random quantities, where the randomness is due to the randomness in the training data $\mathcal{T}$.

   (a) Show that

$$\mathbb{E}(\text{Err}_{\text{in}}) = n^{-1}\sum_{i=1}^{n}\left(\left[\text{Bias}\left\{\hat{f}(\boldsymbol{x}_i)\right\}\right]^2 + \text{Var}\left\{\hat{f}(\boldsymbol{x}_i)\right\}\right) + \sigma^2,$$

   where the expectation is taken over the training set $\mathcal{T}$.

   (b) Show that the expected **optimism** in the training error, $\omega = \text{Err}_{\text{in}} - \overline{\text{err}}$, is

$$\mathbb{E}(\omega) = \frac{2}{n}\sum_{i=1}^{n}\text{Cov}\left\{y_i, \hat{f}(\boldsymbol{x}_i)\right\},$$

   where the expectation is taken over the training set $\mathcal{T}$. Why does this show that the training error usually **underestimates** the test (i.e. prediction) error?

2. (Ridge regression) Suppose $\boldsymbol{X} = (\boldsymbol{x}_1 \cdots \boldsymbol{x}_n)^T \in \mathbb{R}^{n\times p}$, $\boldsymbol{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathbb{R}^n$ and $\lambda \geq 0$.

   (a) Show that $\boldsymbol{X}^T\boldsymbol{X} + \lambda I_p$ is invertible for all $\lambda > 0$, regardless of whether or not $\boldsymbol{X}$ is full rank.

   (b) **Bonus:** Suppose $\text{Var}(y_i) = \sigma^2$. Recall that because ridge regression is linear in $\boldsymbol{Y}$, $df_\lambda = \frac{1}{\sigma^2}\text{Tr}(\boldsymbol{H}_\lambda)$, where $\hat{\boldsymbol{Y}}_\lambda^{(ridge)} = \boldsymbol{X}\hat{\boldsymbol{\beta}}^{(ridge)}(\lambda) = \boldsymbol{H}_\lambda\boldsymbol{Y}$ (you should be able to write down $\boldsymbol{H}_\lambda$ in terms of $\boldsymbol{X}$ and λ). Show that for $\lambda_1 < \lambda_2$, $df_{\lambda_1} > df_{\lambda_2}$. That is, increasing λ reduces the effective number of parameters in the model.

3. Consider the data Fat.txt. Remove every tenth observation from the data for use as a test sample. Use the remaining data to fit (i.e. train) the following models where % body fat, `siri`, is the response and all other variables are predictors:

(i) A simple linear model.

(ii) Ridge regression, where the tuning parameter λ is chosen with generalized cross validation (see the lecture slides from 10/22 and 2(b) above). See RidgeExample.R for an example of how to do this in R. Plot the generalized cross validation value as a function of λ so that the minimum value is clearly visible.

(a) Which model has the smaller training error (see problem 1)? Why is training error a poor judge of how well the model will predict future data?

(b) Now use the models you fit in (i) and (ii) to predict the held out data. Which model performs better? Clearly indicate the metric (i.e. loss function) you used to judge model performance. (**Hint**: since you are using squared loss to choose λ, you should be using squared loss to judge prediction...)