Introduction to Statistical Learning Homework 1

Due: 2021.03.09

Lecturer: Prof. Sheng Yu

1 Problem 1

1.1

In OLS, we introduced three sums of squares:

$$SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2 ,$$
$$SS_{reg} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 ,$$
$$SS_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Show that $SS_{tot} = SS_{reg} + SS_{res}$.

1.2

Explain the meaning of R^2 intuitively.

1.3

See **Slides Page 20, Properties of** $\hat{\beta}$, show that $Var(\hat{\beta}_j) = \frac{\sigma^2}{(n-1)S_{x_j}^2} \cdot \frac{1}{1-R_j^2}$.

2 Problem 2

Show that $\hat{\beta}_0^{ols}$ is the Best Linear Unbiased Estimator (BLUE) of β_0 (The intercept of a linear model). *Hint*: First prove it is unbiased, then show its variance is minimal.

3 Problem 3

Recall that in ridge regression:

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_i (Y_i - X_i^T\beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \arg\min_{\beta} (X\beta - Y)^T(X\beta - Y) + \lambda\|\beta\|^2. \tag{1}$$

We have seen in class that the solution of the above formulation is $\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^TY$. In this problem, you will study the interpretation of regularization.

3.1

Instead of viewing β as an unknown deterministic parameter, we can consider β as a random variable whose value is unknown. In this setting, we specify a prior distribution $P(\beta)$ on β that expresses our prior beliefs over the parameters. When data are observed, we can update the beliefs using posterior distribution. In other words, both the prior beliefs on β and the data observation affect the estimation of β. In this fashion, we estimate β using the MAP (maximum a posteriori) estimate as:

$$\hat{\beta}^{MAP} = \arg\max_{\beta} \prod_{i=1}^{n} P(Y_i|X_i;\beta)P(\beta). \tag{2}$$

Show that maximizing Equation 2 can be expressed as minimizing Equation 1 given a Gaussian prior on β (i.e. $P(\beta) \sim \mathcal{N}(0, I\sigma^2/\lambda)$). That is, show that the L2-norm regularization in the linear regression model is effectively imposing a Gaussian prior assumption on the unknown parameter β.

3.2

What is the probabilistic interpretation if $\lambda \to 0$? How about if $\lambda \to \infty$? *Hint*: Consider how the prior $P(\beta) \sim \mathcal{N}(0, I\sigma^2/\lambda)$ is affected by changing λ.

4 Problem 4

Write an **R** or **Python** function that performs K-fold cross-validation procedure to tune the penalty parameter λ in ridge regression using the prostate cancer data: Plot training error, test error, and 5-fold and 10-fold cross-validation errors on the same plot for each value in the sequence of λ that you choose. What is the value of λ proposed by your cross-validation procedure? Comment on the shapes of the error curves.
Hint: the outcome in the prostate cancer data is lpsa.