STAT 435
SPRING QUARTER 2018

Homework # 4
Due Friday, May 4, 2018 at 12:00 PM (Noon)
Online Submission Via Canvas

Instructions: You may discuss the homework problems in small groups, but you must write up the final solutions and code yourself. Please turn in your code for the problems that involve coding. However, for the problems that involve coding, you must also provide written answers: you will receive no credit if you submit code without written answers. You might want to use Rmarkdown to prepare your assignment.

1. Consider the validation set approach, with a 50/50 split into training and validation sets:

   (a) Suppose you perform the validation set approach twice, each time with a different random seed. What's the probability that an observation, chosen at random, is in *both* of those training sets?

   (b) If you perform the validation set approach repeatedly, will you get the same result each time? Explain your answer.

2. Consider K-fold cross-validation:

   (a) Consider the observations in the 1st fold's training set, and the observations in the 2nd fold's training set. What's the probability that an observation, chosen at random, is in *both* of those training sets?

   (b) If you perform K-fold CV repeatedly, will you get the same result each time? Explain your answer.

3. Now consider leave-one-out cross-validation:

   (a) Consider the observations in the 1st fold's training set, and the observations in the 2nd fold's training set. What's the probability that an observation, chosen at random, is in *both* of those training sets?

   (b) If you perform leave-one-out cross-validation repeatedly, will you get the same result each time? Explain your answer.

4. Consider a very simple model,

$$Y = \beta + \epsilon,$$

where Y is a scalar response variable, $\beta \in \mathbb{R}$ is an unknown parameter, and ϵ is a noise term with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$. Our goal is to estimate β. Assume that we have n observations with uncorrelated errors.

   (a) Suppose that we perform least squares regression using all n observations. Prove that the least squares estimator, $\hat{\beta}$, equals $\frac{1}{n}\sum_{i=1}^{n} Y_i$.

   (b) Suppose that we perform least squares using all n observations. Prove that the least squares estimator, $\hat{\beta}$, has variance σ^2/n.

   (c) Consider the least squares estimator of β fit using just $n/2$ observations. What is the variance of this estimator?

   (d) Consider the least squares estimator of β fit using $n(K-1)/K$ observations, for some $K > 2$. What is the variance of this estimator?

   (e) Consider the least squares estimator of β fit using $n-1$ observations. What is the variance of this estimator?

   (f) Derive an expression for $E(\hat{\beta})$, where $\hat{\beta}$ is the least squares estimator fit using all n observations.

   (g) Using your results from the earlier sections of this question, argue that the validation set approach tends to *over*-estimate the expected test error.

   (h) Using your results from the earlier sections of this question, argue that leave-one-out cross-validation does not substantially over-estimate the expected test error, provided that n is large.

   (i) Using your results from the earlier sections of this question, argue that K-fold CV provides an over-estimate of the expected test error that is somewhere between the big over-estimate resulting from the validation set approach and the very mild over-estimate resulting from leave-one-out CV.

5. As in the previous problem, assume

$$Y = \beta + \epsilon,$$

where Y is a scalar response variable, $\beta \in \mathbb{R}$ is an unknown parameter, and ϵ is a noise term with $E(\epsilon) = 0$, $Var(\epsilon) = \sigma^2$. Our goal is to estimate β. Assume that we have n observations with uncorrelated errors.

   (a) Suppose that we perform K-fold cross-validation. What is the correlation between $\hat{\beta}^1$, the least squares estimator of β that we obtain from the 1st fold, and $\hat{\beta}^2$, the least squares estimator of β that we obtain from the 2nd fold?

(b) Suppose that we perform the validation set approach twice, each time using a different random seed. Assume further that exactly $0.25n$ observations overlap between the two training sets. What is the correlation between $\hat{\beta}^1$, the least squares estimator of β that we obtain the first time that we perform the validation set approach, and $\hat{\beta}^2$, the least squares estimator of β that we obtain the second time that we perform the validation set approach?

(c) Now suppose that we perform leave-one-out cross-validation. What is the correlation between $\hat{\beta}^1$, the least squares estimator of $\hat{\beta}$ that we obtain from the 1st fold, and $\hat{\beta}^2$, the least squares estimator of β that we obtain from the 2nd fold?

Remark 1: Problem 5 indicates that the $\hat{\beta}$'s that you estimate using LOOCV are very correlated with each other.

Remark 2: You might remember from an earlier stats class that if $X_1, \ldots, X_n$ are uncorrelated with variance σ^2 and mean μ, then the variance of $\frac{1}{n}\sum_{i=1}^{n} X_i$ equals σ^2/n. But if $Cor(X_i, X_k) = \sigma^2$, then the variance of $\frac{1}{n}\sum_{i=1}^{n} X_i$ is quite a bit higher.

Remark 3: Together, problems 4 and 5 might give you some intuition for the following: LOOCV results in an approximately unbiased estimator of expected test error (if n is large), but this estimator has high variance. In contrast, K-fold CV results in an estimator of expected test error that has higher bias, but lower variance.