STAT 435 HW 4

Chongyi Xu

May 3, 2018

Question 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 timme with a different random seed. What is the probability that an observation, chosen at random, is in both of those training sets?

$$\begin{aligned} P(\text{in training set 2}|\text{in training set 1}) &= P(\text{in training set 2}) \cdot P(\text{in training set 1}) \\ &= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \end{aligned}$$

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

The result will be different each time. The validation estimate can be variable, depending on precisely which observations are included in the training set and which observations are included in the validation set (`Textbook p178`).

Question 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?

$$\begin{aligned} P(\text{in 2nd fold training set}|\text{in 1st fold}) &= P(\text{in 2nd fold}) \cdot P(\text{in 1st fold}) \\ &= \frac{k-1}{k} \cdot \frac{k-1}{k} = \frac{(k-1)^2}{k^2} \end{aligned}$$

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

The result will be different each time. Since K-fold CV randomly split the n observations into K non-overlapping groups. And each time we perform the validation, we treat one of the sets as validation set and the rest as training set. By doing this, we will get a different MSE estimate each time since we are basicly using different data as training and testing set.

Question 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?

Assuming there are n distinct observations.

$$\begin{aligned} P(\text{in 2nd fold training set}|\text{in 1st fold}) &= P(\text{in 2nd fold}) \cdot P(\text{in 1st fold}) \\ &= \frac{n-1}{n} \cdot \frac{n-1}{n} = \frac{(n-1)^2}{n^2} \end{aligned}$$

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

The result will be the same. Since we get n distinct fold and each observation will be a validation set with the remaining observations to be training set. And due to that, each time we perform LOOCV, the training set and the test set will be the same, which will give the same result.

Question 4

Consider a very simple model,

$$Y = \beta + \epsilon$$

where Y is a scalar resposne 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 square estimator, $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} Y_i$

Since we are minimizing the square error, and $E(\epsilon) = 0$

$$\begin{aligned}
\frac{d}{d\hat{\beta}} \sum_{i=1}^{n} (Y_i - (\hat{\beta} + \epsilon))^2 &= 0 \\
\frac{d}{d\hat{\beta}} \sum_{i=1}^{n} (Y_i^2 - 2Y_i\hat{\beta} + \hat{\beta}^2) &= 0 \\
-2\sum_{i=1}^{n} Y_i + 2\sum_{i=1}^{n} \hat{\beta} &= 0 \\
\sum_{i=1}^{n} Y_i &= n\hat{\beta} \\
\Rightarrow \hat{\beta} &= \frac{1}{n}\sum_{i=1}^{n} Y_i
\end{aligned}$$

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

First, let's consider $Var(\epsilon) = \sigma^2$

$$\begin{aligned}
Var(\epsilon) &= E[(\epsilon - E[\epsilon])^2] \\
&= E[(Y_i - E[Y_i])^2] \\
&= Var(Y_i) = \sigma^2
\end{aligned}$$

Then, we have our $Var(\hat{\beta})$ to be

$$\begin{aligned}
Var(\hat{\beta}) &= Var(\frac{1}{n}\sum Y_i) \\
&= \frac{1}{n^2}\sum Var(Y_i) \\
&= \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}
\end{aligned}$$

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

$$Var(\beta) = \frac{2\sigma^2}{n}$$

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

$$Var(\beta) = \frac{K\sigma^2}{n(K-1)}$$

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

$$Var(\beta) = \frac{\sigma^2}{n-1}$$

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

$$\begin{aligned} E(\hat{\beta}) &= E(\frac{1}{n}\sum_{i=1}^{n} y_i) \\ &= \frac{1}{n}E(n\bar{y}) \\ &= \bar{y} \end{aligned}$$

(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.

Since we are only using a subset of the observations, say, first half to be the training set and the reminder to be the validation set. Then, we will get our MSE to be $2\sigma^2/n$. However σ^2/n is the "real" error of the fitting model. So the validation set approach over-estimates the test error.

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

Since the LOOCV method only picked one single observation as validation set, the variance, as we calculated in part(e), will be $\frac{\sigma^2}{n-1}$. When n is large, the difference between this error term and σ^2/n is substantially small that we can conclude that this method does not over-estimate the expted test error.

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

From part(d), we have found that for K-fold, the test error will be $\frac{K\sigma^2}{n(K-1)}$. If the K we choose is small, the estimation will be as big over-estimate as the result from validation set approach. However, if the K we choose is large, we will get a estimation close to the result from LOOCV.

Question 5

(a) Suppose that we perform K-fold CV. 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?

Let the first fold has training set to be I and the setcond fold has training set to be J.

$$
\begin{aligned}
Cor(\hat{\beta}^1, \hat{\beta}^2) &= \frac{Cov(\hat{\beta}^1, \hat{\beta}^2)}{\hat{\sigma^1}\hat{\sigma^2}} \\
&= \frac{1}{(\sqrt{\frac{K\sigma^2}{n(K-1)}})^2} \cdot \frac{K^2}{n^2(K-1)^2} \sum_{i \in I} \sum_{j \in J} Cov(Y_i, Y_j)
\end{aligned}
$$

Since for 1st fold and 2nd fold, they have $n(K-2)/K$ training observations overlapped, $\sum_{i \in I} \sum_{j \in J} Cov(Y_i, Y_j) = \frac{n(K-2)\sigma^2}{K}$

$$
\begin{aligned}
Cor(\hat{\beta}^1, \hat{\beta}^2) &= \frac{K}{n(K-1)\sigma^2} \frac{n(K-2)}{K} \sigma^2 \\
&= \frac{K-2}{K-1}
\end{aligned}
$$

(b) Suppose that we perform the validation set approach twice, each time using a different random seed. Assmue further that exactly $0.25n$ observation overlap between the two training sets. What is the correlation between $\hat{\beta}^1$ and $\hat{\beta}^2$?

Same as we did in part(a), but this time we have $0.25n$ overlap instead.

$$
\begin{aligned}
Cor(\hat{\beta}^1, \hat{\beta}^2) &= \frac{1}{(\sqrt{\frac{2\sigma^2}{n}})^2} \frac{1}{n^2} \sum \sum Cov(Y_i, Y_j) \\
&= \frac{1}{2n\sigma^2} \frac{n}{4} \sigma^2 \\
&= \frac{1}{8}
\end{aligned}
$$

(c) Now suppose that we perform LOOCV. What is the correlation between $\hat{\beta}^1$ and $\hat{\beta}^2$?

Similarly, but this time we have $n-2$ observations overlapped.

$$
\begin{aligned}
Cor(\hat{\beta}^1, \hat{\beta}^2) &= \frac{1}{(\sqrt{\frac{\sigma^2}{n-1}})^2} \frac{1}{(n-1)^2} \sum \sum Cov(Y_i, Y_j) \\
&= \frac{1}{(n-1)\sigma^2} \cdot (n-2)\sigma^2 \\
&= \frac{n-2}{n-1}
\end{aligned}
$$