An Introduction to Statistical Learning: Chapter 5

Due on January 15,2016 at 3:10pm

G. James et.al.

Tony Jiang

1. Using basic statistical properties of the variance, as well as single- variable calculus, derive (5.6). In other words, prove that α given by (5.6) does indeed minimize $\text{Var}(\alpha X + (1-\alpha)Y)$.

Answers:

$$\begin{aligned} Var(\alpha X + (1-\alpha)Y) &= \alpha^2 Var(X) + (1-\alpha)^2 Var(Y) + 2\cdot\alpha(1-\alpha)Cov(X,Y) \\ &= \alpha^2\sigma_X^2 + (1-2\alpha+\alpha^2)\sigma_Y^2 + (2\alpha - 2\alpha^2)\sigma_{XY} \\ &= (\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\alpha^2 + (2\sigma_{XY} - 2\sigma_Y^2)\alpha + \sigma_Y^2 \end{aligned}$$

To minimize the above, we take derivative with regard to α:

$$\begin{aligned} \frac{d(Var(\alpha X + (1-\alpha)Y))}{d\alpha} &= \frac{d(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\alpha^2 + (2\sigma_{XY} - 2\sigma_Y^2)\alpha + \sigma_Y^2}{d\alpha} \\ &= 2(\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY})\alpha + (2\sigma_{XY} - 2\sigma_Y^2) \\ &= 0 \\ \alpha &= \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}} \end{aligned}$$

2. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.

(b) What is the probability that the second bootstrap observation is not the jth observation from the original sample?

(c) Argue that the probability that the jth observation is not in the bootstrap sample is $(1 - 1/n)^n$.

(d) When $n = 5$, what is the probability that the jth observation is in the bootstrap sample?

(e) When $n = 100$, what is the probability that the jth observation is in the bootstrap sample?

(f) When $n = 10,000$, what is the probability that the jth observation is in the bootstrap sample?

(g) Create a plot that displays, for each integer value of n from 1 to 100,000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

(h) We will now investigate numerically the probability that a boot- strap sample of size n = 100 contains the jth

observation. Here $j = 4$. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

```
>store=rep(NA, 10000)
> for(i in 1:10000){
store[i]=sum(sample(1:100, rep=TRUE)==4)>0
}
> mean(store)
```

Answers:

(a). $p = 1 - \frac{1}{n}$. There are $n-1$ ways to choose the sample from n samples.
(b). same as (a).
(c). Bootstrapping is sampling with replacement, so each sampling is independent from each other. to ensure n observations are all different from the jth observation, this is equivalent of n independent events of (a) or (b): $\left(1 - \frac{1}{n}\right)^n$
(d). $Pr(n = 5) = 1 - \left(1 - \frac{1}{5}\right)^5 = 1 - 0.328 = 67.2\%$
(e). $Pr(n = 100) = 1 - \left(1 - \frac{1}{100}\right)^{1}00 = 63.4\%$
(f). $Pr(n = 100,000) = 1 - \left(1 - \frac{1}{100,000}\right)^{1}00,000 = 63.2\%$
(g).

```
pr = function(n) return(1 - (1 - 1/n)^n)
x = 1:1e+05
plot(x, pr(x))
```

(h). the result shows a average probability of 64.68% which is very close to theoretical value given in (e).

3. We now review k-fold cross-validation.
(a) Explain how k-fold cross-validation is implemented.
(b) What are the advantages and disadvantages of k-fold cross-validation relative to:
i. The validation set approach?
ii. LOOCV?

Answers:

(a). The set of observations are randomly divided into k groups (folds) of equal size. The first fold is used as test set and the remaining $k-1$ folds are used as training set. The mean squared error is computed for each iteration then averaged across all iterations to give estimate of the MSE.

(b).
i. The validation set approach is conceptually simple and easily implemented as you are simply partitioning the existing training data into two sets. However, there are two drawbacks: (1.) the estimate of the test error rate can be highly variable depending on which observations are included in the training and validation sets. (2.) the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.
ii. LOOCV is a special case of k-fold cross-validation with $k = n$. Thus, LOOCV is the most computationally intense method since the model must be fit n times. Also, LOOCV has higher variance, but lower bias, than k-fold CV.

4. Suppose that we use some statistical learning method to make a prediction for the response Y for a particular value of the predictor X. Carefully describe how we might estimate the standard deviation of our prediction.

Answers:

We can use the bootstrap approach. The bootstrap approach works by repeatedly sampling observations (with replacement) from the original data set n times each time fitting a new model and subsequently obtaining the RMSE of the estimates for all n models. Then standard deviation of the prediction can be approximated by computing the variation in all our estimates from our n model parameters.