Weekly Summary 2

Seline Yang

2/5/2018

1. k-fold Cross Validation

Purpose: k-fold Cross Validation is one of the resampling techniques. The method is useful for estimating the test MSE. It helps us to determine how well our estimate would do on test data.

Summary: k-fold CV is an alternative to LOOCV. Instead of creating n test sets, we create k test sets. The dataset is randomly divided into k groups of approximately equal size. Setting fold 1 to be the validation set, and the remaining as the training data. Repeat the process with remaining k-1 folds. Compute MSE for each process. The two default choices of k is 5 and 10. If it is very important to have low variance, we choose $k = 10$.

Pros: Compare to LOOCV, this method is much more faster to compute, it is an alternative to LOOCV. We will have lower bias compare to the validation method.

Cons: We will have relative higher bias compare to LOOCV in estimating the test MSE. We use fewer data as our training set, and hence the estimate is biased from what we would get if we used all the data to train. We have bias-variance trade-off.

R Commands:

```r
library(ISLR)
library(boot)
set.seed(17)
cv.error.10=rep(0,10)
for (i in 1:10){
glm.fit=glm(mpg~poly(horsepower,i),data=Auto)
cv.error.10[i]=cv.glm(Auto, glm.fit, K=10)$delta[1]
}
cv.error.10
```

```
##  [1] 24.20520 19.18924 19.30662 19.33799 18.87911 19.02103 18.89609
##  [8] 19.71201 18.95140 19.50196
```

2. Bootstrapping

Purpose: To assess the accuracy with which a given method estimates parameters. It is used to estimate the uncertainty associated with given estimate.

Summary: In general circumstances, it is very difficult to estimate the standard error of a statistic. Bootstrapping is a resampling technique that we use to create multiple samples from our one sample. We take a random sample of size n from the training set, sample with replacement. Compute the statistics and repeat the process many times. This will give a large collection of the statistics we need, and we can use them to create the distribution to estimate the variability in the statistic we are interested in and create CI.

Pros: It is very useful to find the distribution of the statistic that we are interested in if we have small sample. It is a straightforward way to derive estimates of standard errors and confidence intervals for complex estimators of complex parameters of the distribution. The statistic we obtained from bootstrap is more accurate than the standard intervals obtained using sample variance and assumptions of normality.

Cons: It the sample distribution is extremely skewed, the bootstrap interval might be unreliable. We need a representative sample from the population. If the sample is biased, the estimates will also be biased.

R Commands:

```
set.seed(2532)
boot = rep(NA,100)
for(i in 1:100){
  temp = sample(StarbucksA$Calories, size = 150,
                replace = TRUE)
  boot[i] = median(temp)
}
```

```
## Error in sample(StarbucksA$Calories, size = 150, replace = TRUE): object 'StarbucksA' not found
```