Summary 1

Seline Yang

1/29/2018

1. Mean Squared Error (MSE)

Purpose: To assess model fit.

Summary: We have $\epsilon = Y - \hat{Y}$, $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$, where n is the number of observation in the training set, $\hat{f}(x_i) = \hat{y_i}$ is the prediction for y_i for observation x_i. We want to minimize the difference between true value and the prediction, so smallest MSE is preferred.

Pros: We can use MSE to compare different models for our prediction, it is also used in stepwise regression to determine how many predictors to include.

Cons: It heavily weight the outliers.

R Commands: mse<-mean(training.model$residuals^2)

2. Validation Method

Purpose: To split the dataset into two subsets: training data and validation data, so that we can use the data set to build prediction models and test our models.

Summary: The size of validation set and trainig set depends on the actual size of the data set we use. This process allow us to obtain an estimate of a test MSE, and hencce assess our model fit.

Pros: Easy way to split the original data set to build model and test the prediction.

Cons: Bias accurs. The results depends in which data points end up in the test set versus the training set. So we might repeat the process more than one time. The test MSE may overestimate the test MSE we would have seen if we has used the entire training set for training.

R Commands:

```
library(ISLR)
data(Auto)
# Count how may training observations we have
n<- nrow(Auto)
# First, create the training set
Training<- sample(1:n, n/2, replace = FALSE) #change 2 to whatever the size you want to split
Training<- Auto[Training, ]
# Everything that is left is in the validation set
#Validation<- Auto[-Training, ]
```

3. Leave-one-out cross-validation (LOOCV)

Purpose: To split the dataset into two subsets: training data and validation data, so that we can use the data set to build prediction models and test our models.

Summary: LOOCV begins by splitting into two datasets, all data points stay in the training set except one. We perform this process multiple times, so that each of the n training points gets left out exactly once.

Pros: Less bias than validation approach. There is less variance in the estimate of the test MSE than in the validation approach. There is no randomness in the training/validation splits. It can be used for logistic regression, linear regression and more.

Cons: The runtime is long, and it will be an issue when n is very large.

R Commands:

```r
# Load the Auto Data
library(ISLR)
data(Auto)
# Count the number of data points in the training data
n = nrow(Auto)
# Fit the model with the linear term only
M1 <- lm( mpg~ horsepower, data = Auto[-1,]  )
# Fit with the linear and the quadratic
M2 <- lm(mpg~ horsepower +I(horsepower^2),data = Auto[-1,]  )
# Create a storage space for our prediction
haty_M1 <- rep(0,n)
haty_M2 <- rep(0,n)

# Run LOOCV
for( i in 1:n){
    # Fit the model with the linear term only
    M1 <- lm( mpg~ horsepower, data = Auto[-i,]  )
    # Fit with the linear and the quadratic
    M2 <- lm(mpg~ horsepower +I(horsepower^2),data = Auto[-i,])
    # Predict for the value that we left out and store
    haty_M1[i] <- predict(M1,Auto[i,])
    haty_M2[i] <- predict(M2,Auto[i,])
}
```