Homework Week 3

Kai Li

2/1/2017

## Question 1: k-fold cross-validation

#### (a) Explain how k-fold cross-validation is implemented

To implement k-fold cross-validation, we need to randomly divide the dataset into k different parts. For each part, MSE is calculated based on the application of a model drawn from the k-1 parts on the k part. And after repeating this step for k different times, the mean MSE is calculated as the evaluation of the model.

#### (b) Advantages and disadvantage of k-fold cross-validation to the validation set approach and LOOCV

As compared to the validation set approach, the advantages of k-fold cross-validation include that it is more stable and it makes use of all the data points.

In its nature, LOOCV is just a special case of k-fold cross-validation. And one obvious advantage of k-fold cross-validation over LOOCV is that the calculation is much lighter. However, on the other side of generating lower variance, k-fold cross-validation could generate higher bias than LOOCV, which is one of the disadvantages of this method.

## Question 2: compute standard errors using bootstrap and glm

#### (a)

data <- Default  
  
set.seed(1)  
model <- glm(default ~ income + balance,   
 data = data,  
 family = binomial(link = "logit"))  
  
summary(model)

##   
## Call:  
## glm(formula = default ~ income + balance, family = binomial(link = "logit"),   
## data = data)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.4725 -0.1444 -0.0574 -0.0211 3.7245   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.154e+01 4.348e-01 -26.545 < 2e-16 \*\*\*  
## income 2.081e-05 4.985e-06 4.174 2.99e-05 \*\*\*  
## balance 5.647e-03 2.274e-04 24.836 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2920.6 on 9999 degrees of freedom  
## Residual deviance: 1579.0 on 9997 degrees of freedom  
## AIC: 1585  
##   
## Number of Fisher Scoring iterations: 8

A logistic regression model is established between the dependent variable "default" and independent variables "income" and "balance" using the dataset "Default". As shown in the results above, the standard errors of the two variables of income and balance are 4.985e-06 and 2.274e-04, respectively.

#### (b) & (c)

boot.fn <- function(data, index) {  
 model <- glm(default ~ income + balance,  
 data[index,],  
 family = binomial(link = "logit"))  
 return(c(summary(model)$coefficients[2, 1],  
 summary(model)$coefficients[3, 1]))  
}  
  
set.seed(1)  
table <- boot(data, boot.fn, 1000)

#### (d)

As is shown in the results above, if seed is set as 1, the standard errors of the following two variables are:

* income: 4.985e-06 using glm model and 4.960e-06 using bootstrapping method
* balance: 2.274e-04 using glm model and 2.313e-04 using bootstrapping method

It seems that the standard errors from both methods are highly identical.