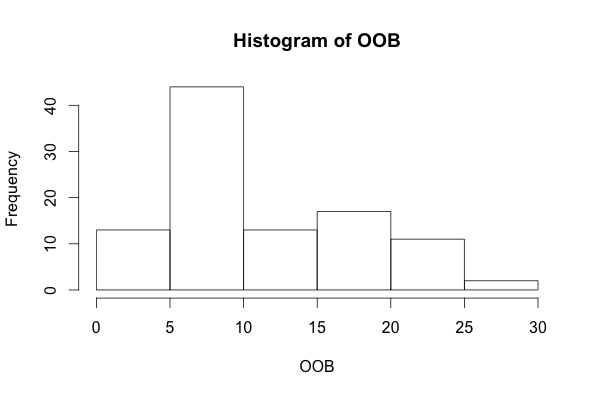
Random Forests and Cross Validation

This is an open-ended homework. The focus on ‘Sepsis’ data.Select 70% of the data at random. This is training data. Run a random forest on it. Run the forest through the remaining data. Note down the OOB. Repeat this exercise 10 times. Get the histogram of OOB. Get a non-parametric density estimate based on OOBs. Get a 95% confidence interval of the population OOB.

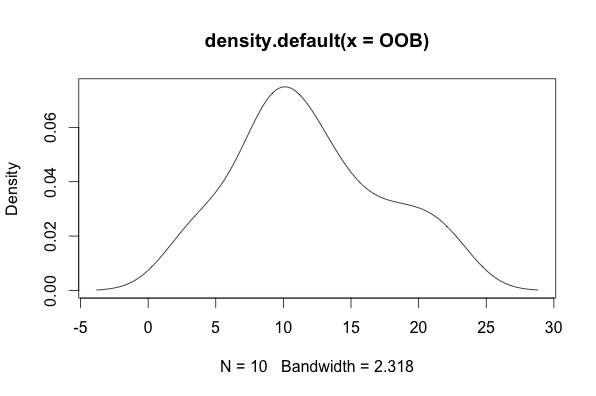
OOBs for 10 iterations are:

9.375 18.750 18.750 9.375 9.375 21.875 6.250 15.625 12.500 15.625

Histogram of OOB



Non-parametric density estimate



Call:

density.default(x = OOB)

Data: OOB (10 obs.); Bandwidth 'bw' = 2.922

x y

Min. :-2.516 Min. :0.0001604

1st Qu.: 5.773 1st Qu.:0.0049340

Median :14.062 Median :0.0305752

Mean :14.062 Mean :0.0301223

3rd Qu.:22.352 3rd Qu.:0.0553148

Max. :30.641 Max. :0.0601473

The code:

library(randomForest)

getwd()

setwd("/Users/aishwaryakulkarni/Documents/Comp\_Stats")

Sepsis<-read.table("Sepsis.txt",header=T)

head(Sepsis)

dim(Sepsis)

table(Sepsis$Death)

Sepsis$Death<-as.factor(Sepsis$Death)

Training<-matrix(0,74,10)

Index<-1:10

for (i in Index)

{

Training[,i]<-sample(1:106,74)

}

Training[1:10,1:10]

MB<-randomForest(Death ~ .,data=Sepsis[Training[,1],])

MB1<-predict(MB,newdata=Sepsis[-Training[,1],],type="class")

MB1

MB2<-sum(!MB1==Sepsis[-Training[,1],]$Death)

MB2

100\*MB2/32

I<-1:100

OOB<-rep(0,10)

for(i in 1:10)

{

OOB[i] <- 100\*sum(!predict(randomForest(Death~.,data=Sepsis[Training[,i],]),newdata=Sepsis[-Training[,i],],type="class")==Sepsis[-Training[,i],]$Death)/32

}

OOB

summary(OOB)

quantile(OOB,c(0.025,0.975))

mean(OOB)

density(OOB)

d<-density(OOB)

plot(d)

> library(randomForest)

> getwd()

[1] "/Users/aishwaryakulkarni/Documents/Comp\_Stats"

> setwd("/Users/aishwaryakulkarni/Documents/Comp\_Stats")

> Sepsis<-read.table("Sepsis.txt",header=T)

> head(Sepsis)

ID Shock Malnutrition Alcoholism Age Infarction Death

1 1 0 0 0 56 0 0

2 2 0 0 0 80 0 0

3 3 0 0 0 61 0 0

4 4 0 0 0 26 0 0

5 5 0 0 0 53 0 0

6 6 0 1 0 87 0 1

> dim(Sepsis)

[1] 106 7

> table(Sepsis$Death)

0 1

85 21

> Sepsis$Death<-as.factor(Sepsis$Death)

> Training<-matrix(0,74,10)

> Index<-1:10

> for (i in Index)

+ {

+ Training[,i]<-sample(1:106,74)

+ }

> Training[1:10,1:10]

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 43 85 90 90 61 35 90 70 55 80

[2,] 100 62 105 48 33 47 95 36 97 71

[3,] 51 98 48 56 15 95 99 63 15 20

[4,] 60 50 106 86 5 93 8 12 69 2

[5,] 54 41 44 58 82 3 53 14 51 70

[6,] 61 3 42 96 68 19 79 10 20 105

[7,] 4 1 82 39 11 29 25 85 59 98

[8,] 15 18 62 28 75 80 80 106 4 92

[9,] 22 21 104 45 34 96 102 1 101 68

[10,] 106 79 77 3 24 43 4 53 35 64

> MB<-randomForest(Death ~ .,data=Sepsis[Training[,1],])

> MB1<-predict(MB,newdata=Sepsis[-Training[,1],],type="class")

> MB1

5 6 10 11 17 18 26 31 32 34 41 42 44 49 59 62 63 66 68 71 79 81

0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

82 84 86 89 92 95 96 97 102 104

0 0 0 0 0 0 0 1 0 0

Levels: 0 1

> MB2<-sum(!MB1==Sepsis[-Training[,1],]$Death)

> MB2

[1] 3

> 100\*MB2/32

[1] 9.375

> I<-1:100

> OOB<-rep(0,10)

> for(i in 1:10)

+ {

+ OOB[i] <- 100\*sum(!predict(randomForest(Death~.,data=Sepsis[Training[,i],]),newdata=Sepsis[-Training[,i],],type="class")==Sepsis[-Training[,i],]$Death)/32

+ }

> OOB

[1] 9.375 18.750 18.750 9.375 9.375 21.875 6.250 15.625 12.500 15.625

> summary(OOB)

Min. 1st Qu. Median Mean 3rd Qu. Max.

6.250 9.375 14.060 13.750 17.970 21.880

> quantile(OOB,c(0.025,0.975))

2.5% 97.5%

6.953125 21.171875

> mean(OOB)

[1] 13.75

> d<-density(OOB)

> plot(d)

> density(OOB)

Call:

density.default(x = OOB)

Data: OOB (10 obs.); Bandwidth 'bw' = 2.922

x y

Min. :-2.516 Min. :0.0001604

1st Qu.: 5.773 1st Qu.:0.0049340

Median :14.062 Median :0.0305752

Mean :14.062 Mean :0.0301223

3rd Qu.:22.352 3rd Qu.:0.0553148

Max. :30.641 Max. :0.0601473