Mingjie Zhao HW12

1)

[1] 0.07538126

Codes

amp = read.table("dna.txt",sep=" ")

amp = amp[,-394] # The 394th column is completely missing for everybody

tumor = read.delim("label.txt",sep="\t",header=F)

tumor = I(tumor[,2] == "Breast cancer")

tumor = ifelse(is.na(tumor),F,tumor)

totalamp = apply(amp,2,sum)

amp2 = amp[,totalamp >= 100]

# LDA

library(MASS)

amp3=cbind(amp2, tumor)

n=ncol(amp3)

row=sample(c(1:nrow(amp3)),nrow(amp3)/2,replace=FALSE)

train=amp3[row,]

test=amp3[-row,]

ldairis=lda(tumor~.,data=train)

testclass=predict(ldairis,test[,-n])$class

ME=1-sum(test[,n]==testclass)/nrow(test)

3) The answer is no. This may cause “overoptimism” , give lower value than the true value of ME and make the estimate of the ME is biased. This is because pick the number of locations with the smallest error make the training data dependent with the number of locations (you have the preference). Solution is get an independent dataset relative to the collected data, you randomly pick the number of locations and fit LDA, or maybe do crosee-validation.