breastCancer1.r

setwd("C:/Users/Tony/Dropbox/Rowan/DM2/Lecture1/WisconsinBreastCancer")  
f=read.csv(file="wdbc\_data\_abclean.csv",header=F,stringsAsFactors = TRUE)

dim(f)

## [1] 569 32

f[1,]

## V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11  
## 1 842302 1 17.99 10.38 122.8 1001 0.1184 0.2776 0.3001 0.1471 0.2419  
## V12 V13 V14 V15 V16 V17 V18 V19 V20  
## 1 0.07871 1.095 0.9053 8.589 153.4 0.006399 0.04904 0.05373 0.01587  
## V21 V22 V23 V24 V25 V26 V27 V28 V29 V30  
## 1 0.03003 0.006193 25.38 17.33 184.6 2019 0.1622 0.6656 0.7119 0.2654  
## V31 V32  
## 1 0.4601 0.1189

library(plyr)  
count(f,c("V2"))

## V2 freq  
## 1 0 357  
## 2 1 212

## 357 benign, 212 malignant

set.seed(2)  
train=sample(1:nrow(f),nrow(f)\*(8/10))  
test=-train  
  
trainingData=f[train,]  
testingData=f[test,]  
dim(trainingData)

## [1] 455 32

dim(testingData)

## [1] 114 32

library("neuralnet")  
names(trainingData)

## [1] "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8" "V9" "V10" "V11"  
## [12] "V12" "V13" "V14" "V15" "V16" "V17" "V18" "V19" "V20" "V21" "V22"  
## [23] "V23" "V24" "V25" "V26" "V27" "V28" "V29" "V30" "V31" "V32"

formula = V2~V3+V4+V5+V6+V7+V8+V9+V10+V11+V12+V13+V14+V15+V16+V17+V18+V19+V20+V21+V22+V23+V24+V25+V26+V27+V28+V29+V30+V31+V32  
nnet<-neuralnet(formula,trainingData, hidden=64, threshold=0.1)  
results<-compute(nnet,testingData[,3:32])  
testingData$result<- sapply(results$net.result, function(b) {  
 if (b<=.5){  
 return(0)  
 }else{  
 return(1)  
 }})  
count(testingData,c("V2","result"))

## V2 result freq  
## 1 0 0 77  
## 2 0 1 1  
## 3 1 0 4  
## 4 1 1 32

err = 5/(5+77+32)  
err

## [1] 0.04386

## 4% error can probably be improved by tuning hidden layers and threshold

## this version ran in about 30 seconds