Sample ANN Practice Exercise - Drug Reaction

PVS

Wednesday, February 22, 2017

## Problem statement

A Pharmaceutical firm that developed a particular drug for women wants to understand the characteristics that cause some of them to have an adverse reaction to a particular drug. They collect data on 15 women who had such a reaction and 15 women who did not. The variable measured are \* Systolic Blood Pressure \* Cholesterol level \* Age of the person \* Whether or not the woman was pregnant (1=yes) \* The dependent variable indicates if there was an adverse reaction (1=yes)

Using neural network classify and predict five cases.

## Solution

The sample data size is 30; we split that into training data and test data in the ration 25:5

## Load required packages

library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(nnet)# limited plot methods but very good to start learning ANN,Can use Neuralnet library  
## check "https://beckmw.wordpress.com/tag/nnet/" to learn more  
library(pROC)

## Type 'citation("pROC")' for a citation.

##   
## Attaching package: 'pROC'

## The following objects are masked from 'package:stats':  
##   
## cov, smooth, var

library(e1071)

## set working directory and read input data

setwd ("D:/NN")  
getwd()

## [1] "D:/NN"

data\_nn = read.table("D:/NN/DrugReaction.csv",sep = ",", header = T)

## check the data distribution & type

str(data\_nn)

## 'data.frame': 30 obs. of 5 variables:  
## $ BP : int 100 120 110 100 95 110 120 150 160 125 ...  
## $ Cholesterol : int 150 160 150 175 250 200 180 175 185 195 ...  
## $ Age : int 20 16 18 25 36 56 59 45 40 20 ...  
## $ Pregnant : int 0 0 0 0 0 0 0 0 0 1 ...  
## $ Drug.Reaction: int 0 0 0 0 0 0 0 0 0 0 ...

# Define a function to find the type of variable in the data frame, hrData  
  
findClass <- function (x) {  
 for (i in 1:length(x))  
 {  
 l1 <- names(x[i])  
 l2 <- class(x[[i]])  
 cat("\n ",l1,": ",l2)  
 }   
}  
  
#  
cat("\n Variables according to the class \n")

##   
## Variables according to the class

findClass(data\_nn) ## Get the list of type of variables

##   
## BP : integer  
## Cholesterol : integer  
## Age : integer  
## Pregnant : integer  
## Drug.Reaction : integer

**Observation**

There are 5 variables and 30 observations.

**A) Numerical variables:**

1. BP
2. Cholesterol
3. Age

**B) Integer variables -categorical**

1. Pregnant
2. Drug.Reaction

## Data cleaning

#### 1 Find if there are any missing values

#### 2 Remove Feature with constant values else will give error for Neural Network

#### 3 Remove columns with near zero variance as features should has variance in its distribution.

### 1 Find if there are any missing values  
  
cat("\n Variables with number of missing values \n")

##   
## Variables with number of missing values

sapply(data\_nn, function(x) sum(is.na(x))) # To report missing values

## BP Cholesterol Age Pregnant Drug.Reaction   
## 0 0 0 0 0

### 2 Find if there are any constant values for variables  
  
cat("\n Variables with constant values \n")

##   
## Variables with constant values

sapply(data\_nn, function(x) length(unique(x)))

## BP Cholesterol Age Pregnant Drug.Reaction   
## 15 14 22 2 2

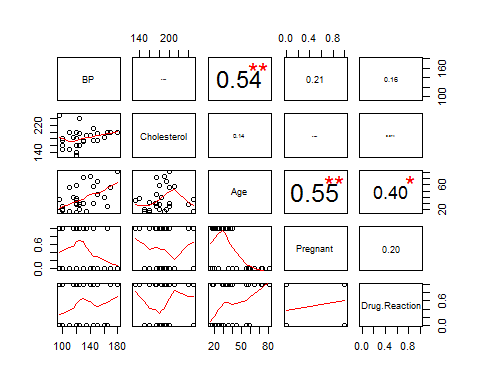
### 3 Find if any variables have near zero variance  
  
nzv <- nearZeroVar(data\_nn)  
nzv

## integer(0)

**Observation**

* There are **No variables** with missing values or constant values or near zero variance

## Exploratory Data Analysis

 **Observation**

1. We find there is no linear relationship between pair of variables except for

* BP and Age - with correlation coefficient is 0.54
* Age and Pregnant - with correlation coefficient is 0.55
* Age and Drug Reaction - with correlation coefficient is 0.40

1. Correlation coefficient is also very low for these variables - the highest being 0.20 (r for Drug.Reaction and Pregnant)

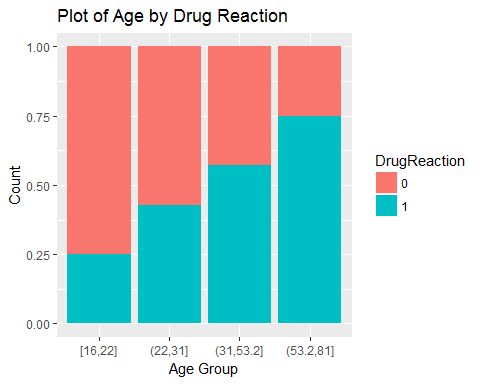
### Visual Representation

#### Drug Reaction % by Age

##   
## Attaching package: 'gtools'

## The following object is masked from 'package:e1071':  
##   
## permutations

##   
## Drug Reaction % by Age



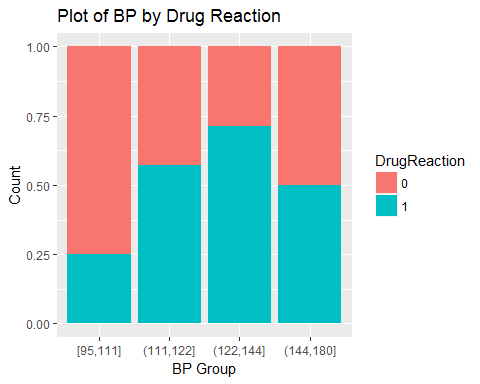
##   
##   
## In the age group [16,22], 75 % of participants have no Drug Reaction and   
## 25 % of participants have Drug Reaction

##   
##   
## In the age group [22,31], 57.14 % of participants have no Drug Reaction and   
## 42.86 % of participants have Drug Reaction

##   
##   
## In the age group [31,53.2], 42.86 % of participants have no Drug Reaction and   
## 57.14 % of participants have Drug Reaction

##   
##   
## In the age group [53.2,81], 25 % of participants have no Drug Reaction and   
## 75 % of participants have Drug Reaction

##   
## Drug Reaction % by BP



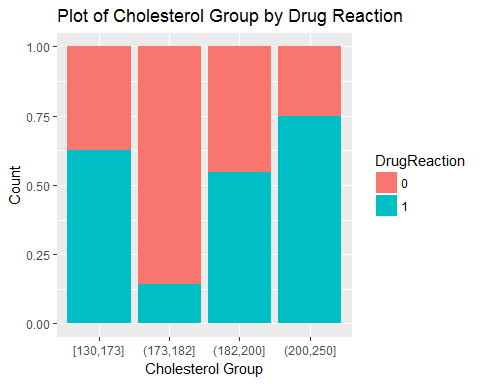
##   
##   
## In the BP group [95,111], 75 % of participants have no Drug Reaction and   
## 25 % of participants have Drug Reaction

##   
##   
## In the BP group [111,122], 42.86 % of participants have no Drug Reaction and   
## 57.14 % of participants have Drug Reaction

##   
##   
## In the BP group [122,144], 28.57 % of participants have no Drug Reaction and   
## 71.43 % of participants have Drug Reaction

##   
##   
## In the BP group [144,180], 50 % of participants have no Drug Reaction and   
## 50 % of participants have Drug Reaction

##   
## Drug Reaction % by Cholesterol Group



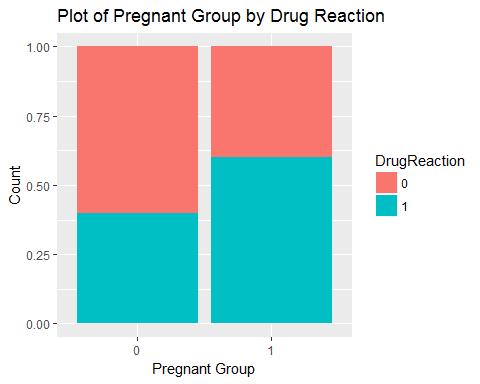
##   
##   
## In the Cholesterol Group [130,173], 37.5 % of participants have no Drug Reaction and   
## 62.5 % of participants have Drug Reaction

##   
##   
## In the Cholesterol Group[173,182], 85.71 % of participants have no Drug Reaction and   
## 14.29 % of participants have Drug Reaction

##   
##   
## In the Cholesterol Group [182,200], 45.45 % of participants have no Drug Reaction and   
## 54.55 % of participants have Drug Reaction

##   
##   
## In the Cholesterol Group [200,250], 25 % of participants have no Drug Reaction and   
## 75 % of participants have Drug Reaction

##   
## Drug Reaction % by Pregnant Group



##   
##   
## In the Non-Pregnant Group, 60 % of participants have no Drug Reaction and   
## 40 % of participants have Drug Reaction

##   
##   
## In the Pregnant Group, 40 % of participants have no Drug Reaction and   
## 60 % of participants have Drug Reaction

## Data Distribution

### Understand dependent variable distribution split

reaction.ratio = table(data\_nn$Drug.Reaction)  
reaction.ratio

##   
## 0 1   
## 15 15

**Observation**

|  |  |  |
| --- | --- | --- |
| SNo | Description | Observation count |
| 1 | Observations with no Drug Reaction | 15 |
| 2 | Observations with Drug Reaction | 15 |

### understand Distribution in percentage terms

reaction.prop = prop.table(reaction.ratio)  
reaction.prop

##   
## 0 1   
## 0.5 0.5

|  |  |  |
| --- | --- | --- |
| SNo | Description | Percentage |
| 1 | Observations with no Drug Reaction | 50% |
| 2 | Observations with Drug Reaction | 50% |

## 

## Split data into Training and Test dataset

set.seed(100)  
n <- 30; trainIndex = sample(1:n, size = round(5/6 \* n), replace=FALSE)  
train.data <- data\_nn[trainIndex,]  
test.data <- data\_nn[-trainIndex,]

## Check if distribution of partition data is correct

prop.table((table(train.data$Drug.Reaction)))

##   
## 0 1   
## 0.52 0.48

**Observations**

#### Distribution of Drug Reaction cases in training data set

|  |  |  |
| --- | --- | --- |
| SNo | Description | Percentage |
| 1 | Observations with no Drug Reaction | 52% |
| 2 | Observations with Drug Reaction | 48% |

prop.table((table(test.data$Drug.Reaction)))

##   
## 0 1   
## 0.4 0.6

**Observations**

#### Distribution of Drug Reaction cases in test data set

|  |  |  |
| --- | --- | --- |
| SNo | Description | Percentage |
| 1 | Observations with no Drug Reaction | 40% |
| 2 | Observations with Drug Reaction | 60% |

* So the data is almost well distributed in the training data and test data sets in between 40 - 52% to 60 to 48%

train.data$Drug.Reaction <- as.factor(train.data$Drug.Reaction)  
train.data$Pregnant <- as.factor(train.data$Pregnant)  
  
levels(train.data$Drug.Reaction) <- c("0","1")  
levels(train.data$Pregnant) <- c("0","1")  
levels(train.data$Drug.Reaction) <- make.names(levels(factor(train.data$Drug.Reaction)))  
  
test.data$Drug.Reaction <- as.factor(test.data$Drug.Reaction)  
test.data$Pregnant <- as.factor(test.data$Pregnant)  
  
levels(test.data$Drug.Reaction) <- c("0","1")  
levels(test.data$Pregnant) <- c("0","1")  
levels(test.data$Drug.Reaction) <- make.names(levels(factor(test.data$Drug.Reaction)))

### Create a control structure for traing NN also with required switches

### Set metric value as Accuracy or Kappa for neural Network model  
metric <- "Accuracy"  
  
# Create a control structure for traing NN also with required switches  
  
#Cross Validation  
cctrl <- trainControl(  
 method = 'cv',number = 10,classProbs = TRUE,  
 verboseIter = TRUE, summaryFunction = twoClassSummary,  
 preProcOptions = list(  
 thresh = 0.75,ICAcomp = 3, k = 5  
 )   
)  
  
# define decay and sizes of hidden neuron to train algorithm on   
my.grid <- expand.grid(.decay = c(0.1, 0.001, 0.0001), .size = c(5, 10, 15))  
# Train/create Neural Network Model  
nn\_model <-  
 train( Drug.Reaction ~ ., data = train.data,  
 method = 'nnet',preProcess = c('center', 'scale'),  
 trControl = cctrl,tuneGrid = my.grid )

## Warning in train.default(x, y, weights = w, ...): The metric "Accuracy" was  
## not in the result set. ROC will be used instead.

## + Fold01: decay=1e-01, size= 5   
## # weights: 31  
## initial value 17.734462   
## iter 10 value 12.257474  
## iter 20 value 11.879315  
## iter 30 value 11.841306  
…

…

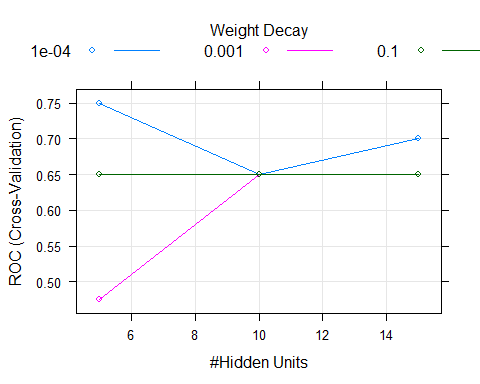
## iter 70 value 0.167361  
## iter 80 value 0.160537  
## iter 90 value 0.153551  
## iter 100 value 0.146776  
## final value 0.146776   
## stopped after 100 iterations

#### view model results

print(nn\_model)

## Neural Network   
##   
## 25 samples  
## 4 predictor  
## 2 classes: 'X0', 'X1'   
##   
## Pre-processing: centered (4), scaled (4)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 23, 22, 23, 23, 23, 23, ...   
## Resampling results across tuning parameters:  
##   
## decay size ROC Sens Spec  
## 1e-04 5 0.750 0.75 0.45  
## 1e-04 10 0.650 0.75 0.50  
## 1e-04 15 0.700 0.75 0.45  
## 1e-03 5 0.475 0.65 0.45  
## 1e-03 10 0.650 0.75 0.50  
## 1e-03 15 0.650 0.75 0.50  
## 1e-01 5 0.650 0.55 0.65  
## 1e-01 10 0.650 0.55 0.65  
## 1e-01 15 0.650 0.55 0.65  
##   
## ROC was used to select the optimal model using the largest value.  
## The final values used for the model were size = 5 and decay = 1e-04.

plot(nn\_model)



#### print variable importance

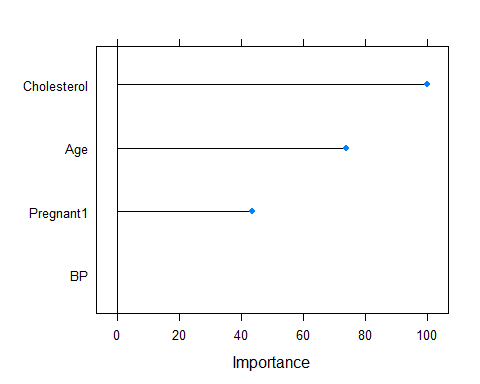
**varImp** is a generic method for calculating variable importance for objects produced by train and method specific methods.

In Neural Network models, the method used is based on Gevrey et al (2003), which uses combinations of the absolute values of the weights. For classification models, the class-specific importances will be the same.

varImp(nn\_model)

## nnet variable importance  
##   
## Overall  
## Cholesterol 100.00  
## Age 73.95  
## Pregnant1 43.60  
## BP 0.00

plot(varImp(nn\_model))



**Observation**

* We observe that the variables are important in classifying Durg.Reaction:

1. Cholesterol
2. Age
3. Pregnant women

### Measuring Model Performance

#### check prediction on Test data

test.pred <- predict(nn\_model, newdata=test.data)  
test.confusion.m <- confusionMatrix(test.pred, test.data$Drug.Reaction)  
test.confusion.m

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction X0 X1  
## X0 2 1  
## X1 0 2  
##   
## Accuracy : 0.8   
## 95% CI : (0.2836, 0.9949)  
## No Information Rate : 0.6   
## P-Value [Acc > NIR] : 0.337   
##   
## Kappa : 0.6154   
## Mcnemar's Test P-Value : 1.000   
##   
## Sensitivity : 1.0000   
## Specificity : 0.6667   
## Pos Pred Value : 0.6667   
## Neg Pred Value : 1.0000   
## Prevalence : 0.4000   
## Detection Rate : 0.4000   
## Detection Prevalence : 0.6000   
## Balanced Accuracy : 0.8333   
##   
## 'Positive' Class : X0   
##

**Observation**

|  |  |
| --- | --- |
| Measure | Value |
| Accuracy of the model | 0.8 |
| Kappa | 0.6154 |
| Sensitivity | 1 |
| Specificity | 0.6667 |
| Hence, the Model performance is good. |  |

### Add scoring data on the test data set

test.data$predict.class <-  
 predict(nn\_model, test.data, type = "raw")  
test.data$predict.score <-  
 predict(nn\_model, test.data, type = "prob")

#### Deciling function to create rank order matrix and to understand the accuracy of the nn model

#### make deciling function

decile <- function(x) {  
 deciles <- vector(length = 10)  
 for (i in seq(0.1, 1, .1)) {  
 deciles[i \* 10] <- quantile(x, i, na.rm = T)  
 }  
   
 return (ifelse(x < deciles[1], 1,  
 ifelse(  
 x < deciles[2], 2,  
 ifelse(x < deciles[3], 3,  
 ifelse(  
 x < deciles[4], 4,  
 ifelse(x < deciles[5], 5,  
 ifelse(  
 x < deciles[6], 6,  
 ifelse(x < deciles[7], 7,  
 ifelse(x < deciles[8], 8,  
 ifelse(x < deciles[9], 9, 10)))  
 ))  
 ))  
 )))  
}

### deciling

test.data$deciles <- decile(test.data$predict.score[, 2])

#### Ranking code

library(data.table)  
# make a copy of final data for any unseen error and calculate KS statistics  
tmp\_DT <- data.table(test.data)  
rank <- tmp\_DT[, list(  
 cnt = length(Drug.Reaction),  
 cnt\_resp = sum(ifelse(Drug.Reaction == 'X1', 1, 0)),  
 cnt\_non\_resp = sum(ifelse(Drug.Reaction == 'X0', 1, 0))  
) ,  
by = deciles][order(-deciles)]  
rank$rrate <- round(rank$cnt\_resp \* 100 / rank$cnt, 2)  
  
rank$cum\_resp <- cumsum(rank$cnt\_resp)  
rank$cum\_non\_resp <- cumsum(rank$cnt\_non\_resp)  
rank$cum\_rel\_resp <- round(rank$cum\_resp / sum(rank$cnt\_resp), 2)  
  
rank$cum\_rel\_non\_resp <-  
 round(rank$cum\_non\_resp / sum(rank$cnt\_non\_resp), 2)  
  
rank$ks <- abs(rank$cum\_rel\_resp - rank$cum\_rel\_non\_resp)  
  
View(rank)

#### 

#### Plot ROC and AUC Curve

library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':  
##   
## lowess

library(pROC)

### Plot ROC and Compute AUC for predicting Class with the model

prob <- predict(nn\_model, newdata = test.data, prob = TRUE)  
  
rocCurve<-roc(response=test.data$Drug.Reaction,predictor=test.data$predict.score[,2]  
 ,levels = rev(levels(test.data$Drug.Reaction)))  
auc(rocCurve)

## Area under the curve: 1

#### Area under the curve: 1

#### Confidence Intervals

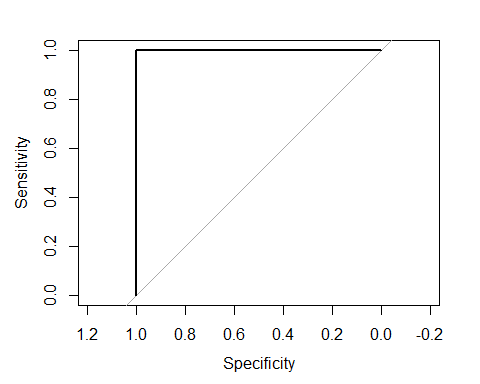
ci.auc(rocCurve)

## Warning in ci.auc.roc(rocCurve): ci.auc() of a ROC curve with AUC == 1 is  
## always 1-1 and can be misleading.

## 95% CI: 1-1 (DeLong)

#### 95% CI: 1 - 1 (DeLong)

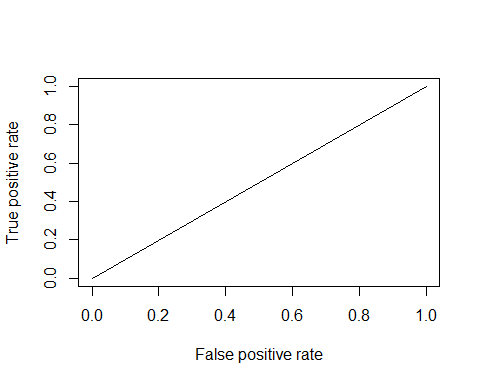
plot.roc(rocCurve)



### Plot True/False positive rates

predvec <- ifelse(prob == "Yes", 1, 0)  
realvec <- ifelse(test.data$Drug.Reaction == "X1", 1, 0)  
pr <- prediction(predvec, realvec)  
  
pref <- performance(pr, "tpr", "fpr")

plot(pref)



KS <- max(attr(pref, 'y.values')[[1]] - attr(pref, 'x.values')[[1]])  
KS

## [1] 0

## [1] 0  
# AUC calculation  
auc <- performance(pr, measure = "auc")  
auc <- auc@y.values[[1]]  
auc

## [1] 0.5

### Observation

Area under the curve is 0.5 indicating model performance is not good.

### Conclusion

### Problem statement

To understand the characteristics that cause some of women to have an adverse reaction to a particular drug

### Conclusion

* We observe that the characteristics that cause some of women to have an adverse reaction to a particular drug are:

1. Cholesterol
2. Age
3. Pregnant women