HW10-Solutions

Problem 1

Refer to the Prostate Cancer data set in Appendix C.5 and Homework 9. Select a random sample of 65 observations to use as the model-building data set (use set.seed(1023)). Use the remaining observations for the test data. (10 pts)

###a) Develop a neural network model for predicting PSA. Justify your choice of number of hidden nodes and interpret your model. Test the model performance on the test data. ###b) Compare the performance of your neuron network model with regression tree model obtained in HW9. Which model is more easily interpreted and why? (5pts) ###c) Compare the performance of your neural network model with that of the best regression model obtained in homework 8. Which model is more easily interpreted and why?

a)

_Solution: We tried single layer NN model and run a simulations with different number of hidden notes. NN with 8 hidden is the best model. The model performance was bad on the hold out samplem, indicating overfit problem on the development sample.

```
library(knitr)
library(neuralnet)
Prostate.Cancer <- read.csv("/cloud/project/Prostate Cancer.csv")
n<-dim(Prostate.Cancer)[1]
set.seed(1023)
sample.ind <- sample(1:n, size = 65)
dev.sample <- Prostate.Cancer[sample.ind,]
holdout.sample <- Prostate.Cancer[-sample.ind,]

pop<-dev.sample
max = apply(pop, 2 , max)
min = apply(pop, 2 , min)
scaled = as.data.frame(scale(pop, center = min, scale = max - min))

max = apply(holdout.sample, 2 , max)
min = apply(holdout.sample, 2 , min)
scaled.holdout.sample = as.data.frame(scale(holdout.sample, center = min, scale = max - min))

### Function for simulation ####
NN.SIM.FUNC<-function(n){
out<-matrix(0,ncol=2,nrow=n)
for (i in 1:n){
NN = neuralnet(PSA.level~Cancer.volume+Weight+Age+Benign.prostatic.hyperplasia+Seminal.vesicle
.invasion+Capsular.penetration+Gleason.score, scaled , hidden = i , linear.output = T)
```

```
predict_testNN = compute(NN, scaled [,c(2:8)])
predict_testNN1 = (predict_testNN$net.result * (max(pop$PSA.level) - min(pop$PSA.level))) + mi
n(pop$PSA.level)
SSE.NN<-sum((pop$PSA.level-predict_testNN1)^2)
out[i,]<-cbind(i,SSE.NN)
}
out
}
NN.SIM.FUNC(10)
```

```
##       [,1]      [,2]
##  [1,]    1  6404.056
##  [2,]    2  7715.066
##  [3,]    3 11341.465
##  [4,]    4  3151.637
##  [5,]    5  4488.283
##  [6,]    6  2516.609
##  [7,]    7  3420.768
##  [8,]    8  2298.254
##  [9,]    9  2839.584
## [10,]   10  3583.955
```

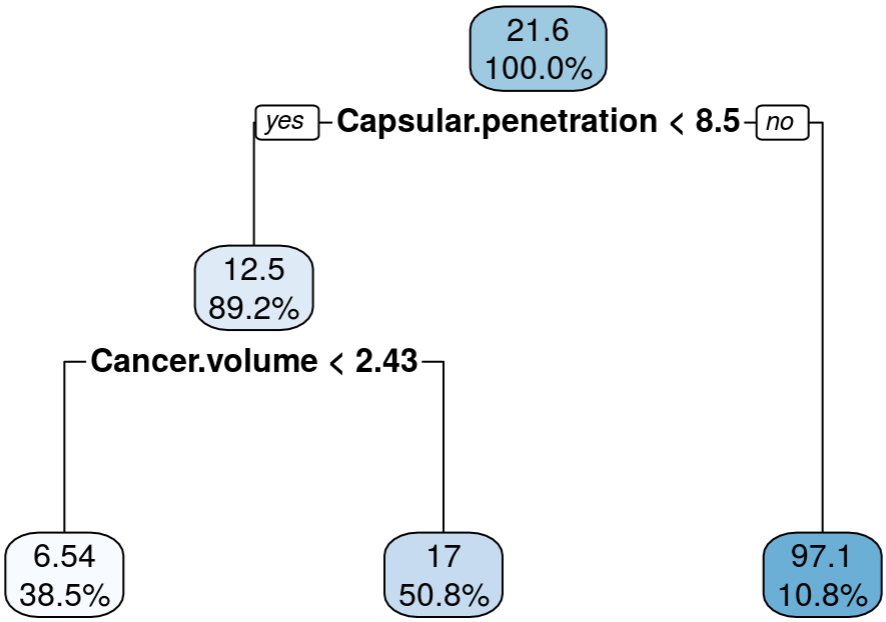
```
#### End Simulations

NN = neuralnet(PSA.level~Cancer.volume+Weight+Age+Benign.prostatic.hyperplasia+Seminal.vesicle
.invasion+Capsular.penetration+Gleason.score, scaled , hidden = 8 , linear.output = T)

predict_testNN = compute(NN, scaled.holdout.sample[,c(2:8)])
predict_testNN1 = (predict_testNN$net.result * (max(holdout.sample$PSA.level) - min(holdout.sa
mple$PSA.level))) + min(holdout.sample$PSA.level)
SSE.NN<-sum((holdout.sample$PSA.level-predict_testNN1)^2)
```

b) On the development sample, NN performs much better than the tree model. However, on the hold out sample, the performance is bad.

```
library(rpart)
library(rpart.plot)
tmod<-rpart(PSA.level~.,dev.sample)
rpart.plot(tmod, digits = 3)
```




```
sse.tree<-sum((predict(tmod)-dev.sample$PSA.level)^2)
sse.tree
```

```
## [1] 54738.96
```

```
sse.tree.holdout<-sum((predict(tmod,holdout.sample)-holdout.sample$PSA.level)^2)
sse.tree.holdout
```

```
## [1] 69339.32
```

c) On the development sample, NN performs much better than the best subset model. However, on the hold out sample, the performance is bad. The best subset model's performance is stable on the development and holdout sample.

```
f.q1.bestsubset<-lm(PSA.level~Cancer.volume+Capsular.penetration,data=Prostate.Cancer)
anova(f.q1.bestsubset)
```

```
## Analysis of Variance Table
##
## Response: PSA.level
```

```
##                       Df Sum Sq Mean Sq F value    Pr(>F)
## Cancer.volume          1  62202   62202  62.757 4.654e-12 ***
## Capsular.penetration   1   4300    4300   4.338   0.03999 *
## Residuals             94  93170     991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ei<-predict(f.q1.bestsubset,holdout.sample)-holdout.sample$PSA.level
SSE.holdout<- sum(ei^2)
SSE.holdout
```

```
## [1] 44111.94
```

Problem 2

2- Refer to the Disease outbreak data set in Appendix C.10. Savings account status is the response variable and age, socioeconomic status, and city sector are the predictor variables.

##a) Fit logistic regression model to predict the saving account status on the predictor variables in first-order terms and interaction terms for. all pairs of predictor variables. State the fitted response function. ##b) Use the likelihood ratio test to determine whether all interaction terms can be dropped from the regression model; use α = .01. State the alternatives, full and reduced models, decision rule, and conclusion. What is the approximate P-value of the test? ##c) Conduct the Hosmer-Lemeshow goodness of fit test for the appropriateness of the logistic regression function by forming five groups of approximately 20 cases each; use α = .05.

a)

_Solution: It is logistic regression function, please see below.

```
Disease.Outbreak <- read.csv("/cloud/project/Disease Outbreak.csv")
Y=Disease.Outbreak$Savings.account.status
X1=Disease.Outbreak$Age
X2=(Disease.Outbreak$Socioeconomic.status==2)*1
X3=(Disease.Outbreak$Socioeconomic.status==3)*1
X4=(Disease.Outbreak$Sector==2)*1
X5=Disease.Outbreak$Disease.status
lmod <- glm(Y ~ (X1+X2+X3+X4+X5)^2, family = binomial)
summary(lmod)
```

```
##
## Call:
## glm(formula = Y ~ (X1 + X2 + X3 + X4 + X5)^2, family = binomial)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3855  -0.8886   0.4118   0.7943   2.0273
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.155908   0.587157   0.266  0.79060
## X1           0.035838   0.021966   1.632  0.10277
## X2          -1.306280   0.817101  -1.599  0.10989
## X3          -2.151271   0.758426  -2.836  0.00456 **
## X4           0.916937   0.781780   1.173  0.24084
## X5          -0.946814   1.062247  -0.891  0.37275
## X1:X2        0.008166   0.029619   0.276  0.78278
## X1:X3        0.002890   0.024113   0.120  0.90461
## X1:X4       -0.021077   0.022438  -0.939  0.34755
## X1:X5        0.021247   0.025814   0.823  0.41045
## X2:X3              NA         NA      NA       NA
## X2:X4       -0.131848   0.880545  -0.150  0.88097
## X2:X5       -0.111640   1.044638  -0.107  0.91489
## X3:X4        0.388653   0.867955   0.448  0.65431
## X3:X5       -0.137603   0.958732  -0.144  0.88587
## X4:X5        0.930980   0.835249   1.115  0.26502
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 270.06  on 195  degrees of freedom
## Residual deviance: 212.84  on 181  degrees of freedom
## AIC: 242.84
##
## Number of Fisher Scoring iterations: 5
```

b)

_Solution: They can be dropped.

```
lmodc <- glm(Y ~ X1+X2+X3+X4+X5, family = binomial)
anova(lmodc,lmod,test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: Y ~ X1 + X2 + X3 + X4 + X5
## Model 2: Y ~ (X1 + X2 + X3 + X4 + X5)^2
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       190     215.36
## 2       181     212.84  9   2.5213   0.9803
```

c)

_Solution:Fit is good.

```
library(ResourceSelection)
```

```
## ResourceSelection 0.3-5   2019-07-22
```

```
hoslem.test(lmod$y,fitted(lmod),g=5)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  lmod$y, fitted(lmod)
## X-squared = 0.85353, df = 3, p-value = 0.8366
```

Problem 3

Refer to the Geriatric study. A researcher in geriatrics designed a prospective study to investigate the effects of two interventions on the frequency of falls. One hundred subjects were randomly assigned to one of the two interventions: education only (X1 = 0) and education plus aerobic exercise training (X1 = 1). Subjects were at least 65 years of age and in reasonably good health. Three variables considered to be important as control variables were gender (X2:0=female;1=male), a balance index (X3). and a strength index (X4). The higher balance index, the more stable is the subject and the higher the strength index, the stronger is the subject. Each subject kept a diary recording the number of falls (Y) during the six months of the study.

##a) Fit the regression model. State the estimated regression coefficients, their estimated standard deviations. and the estimated response function. ##b) Assuming that the fitted model is appropriate, use the likelihood ratio test to determine whether gender (X2) can be dropped from the model: State the full and reduced models. decision rule. and conclusion. What is the P-value of the test ##c)Predicted the number of falls for X1=1, X2=0, X3=45, X4=70.

a)

_Solution: All variables except X2 are significant. Please see below for the poisson regression model.

```
Geriatric.Study <- read.csv("/cloud/project/Geriatric Study.csv")
lpos<-glm(Y~.,data=Geriatric.Study,family="poisson")
summary(lpos)
```

```
##
## Call:
## glm(formula = Y ~ ., family = "poisson", data = Geriatric.Study)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1854  -0.7819  -0.2564   0.5449   2.3626
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.489467   0.336869   1.453  0.14623
## X1          -1.069403   0.133154  -8.031 9.64e-16 ***
## X2          -0.046606   0.119970  -0.388  0.69766
## X3           0.009470   0.002953   3.207  0.00134 **
## X4           0.008566   0.004312   1.986  0.04698 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 199.19  on 99  degrees of freedom
## Residual deviance: 108.79  on 95  degrees of freedom
## AIC: 377.29
##
## Number of Fisher Scoring iterations: 5
```

b)

_Solution: yes, it can be dropped.

```
lposc <- glm(Y ~ X1+X3+X4,data=Geriatric.Study,family="poisson")
anova(lposc,lpos,test="Chi")
```

```
## Analysis of Deviance Table
##
## Model 1: Y ~ X1 + X3 + X4
## Model 2: Y ~ X1 + X2 + X3 + X4
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        96     108.94
## 2        95     108.79  1    0.151   0.6976
```

c)

_Solution: The number of falls is predicted to be 0.4 or 0.

```
dat<-data.frame(cbind(X1=1, X2=0, X3=45, X4=70))
predict(lpos,dat)
```

```
##        1
## 0.445822
```