Assignment 12.1

Name : Y Vasudev

Batch : DA with R , Excel and Tableau

1. Use the given link Data Set.

Answer the below questions:

a. Perform ANOVA test on the discriminant analysis scores of nuclear localization signals of both nuclear

and non-nuclear proteins by class variables (Target).

b. Which class is significantly different from others?

#1. Use the given link Data Set.

yeastdata <- read.table("D:/BIG DATA/DATA ANALYTICS WITH R, EXCEL & TABLEAU/12 NON-LINEAR MODELS/yeast\_data.txt", quote="\"", comment.char="")

View(yeastdata)

dim(yeastdata)

library(reshape)

yeastdata <- rename (yeastdata, c(V1 = "Sequence Name",

V2 = "mcg",

V3 = "gvh",

V4 = "alm",

V5 = "mit",

V6 = "erl",

V7 = "pox",

V8 = "vac",

V9 = "nuc",

V10 = "Class Distribution"))

View(yeastdata)

#Answer the below questions:

#a. Perform ANOVA test on the discriminant analysis scores of nuclear localization signals of both nuclear and non-nuclear proteins by class variables (Target).

results <- aov (yeastdata$nuc ~ yeastdata$`Class Distribution`)

summary(results)

#Since p value is significantly lesser than 0.05 we have strong evidence to reject null hypothesis

#and conclude that there is at least one class significantly different than others.

#b. Which class is significantly different from others?

plot(yeastdata$nuc ~ yeastdata$`Class Distribution`, col = heat.colors(10), xlab = "Class Distribution", ylab = "Score of discriminant analysis" , main = "Boxplot across classes")

#Nuclear class is significantly different from others.

#OR

#In one-way ANOVA test, a significant p-value indicates that some of the group means

#are different, but we don't know which pairs of groups are different. It's possible to perform multiple pairwise-comparison,

#to determine if the mean difference between specific pairs of group are statistically significant.

#As the ANOVA test is significant, we can compute Tukey HSD (Tukey Honest Significant Differences, R function: TukeyHSD())

#for performing multiple pairwise-comparison between the means of groups.

#The function TukeyHSD() takes the fitted ANOVA as an argument.

TukeyHSD(results)

#our output is large

#From above output we got some following groups for which there

#is a significant difference between the means.

#NUC-CYT 0.0744429005 0.05314812 0.095737681 0.0000000

#NUC-EXC 0.1051068931 0.04924622 0.160967571 0.0000001

#NUC-ME1 0.0634965035 0.01319448 0.113798528 0.0026743

#NUC-ME2 0.0861881256 0.03912110 0.133255152 0.0000004

#NUC-ME3 0.0616783217 0.03244028 0.090916364 0.0000000

#NUC-MIT 0.0906947151 0.06521504 0.116174388 0.0000000

#POX-NUC -0.0976783217 -0.17037079 -0.024985855 0.0009144

#VAC-NUC -0.0810116550 -0.14102211 -0.021001196 0.0008397