Lab5

Duc, Raymond, Martin

December 11, 2018

Task 2

```
table(autcon$decision)
```

```
##
##  autism control
##      82      64
```

The autcon data set has 146 obsercations and each has 36 features. Above we can see how the data set is devided in two groups. We can see that the distribution is 56%-44% which is quite balanced

Task 3

a) Cross validation is a method which helps us to approximate the error. The default value in Rosetta function is 10 and this method works in a way that data is devided in this case to 10 groups and each time nine of those are use for training and one for testing. The mean of the obtained errors should be an approximation of the total error.

b) The default reduction method is Johnson. This method is used for finding the shortest path in the graph therefore, to find the rules.

c) The default method for discretization is EqualFrequency method and there are used 3+1 = 4 bins. This method simply creates a partition into n+1 bins with 3 cuts in equal length.

```
autconDefault$quality
```

```
##  Accuracy.Mean Accuracy.Median Accuracy.Std Accuracy.Min Accuracy.Max
##       0.821818             0.8     0.083158     0.733333            1
```

d) The accurancy of the model is shown above. we can see that it is about 80%

```
autconDefault$main[1:3,]
```

```
##            FEATURES DECISION               CUTS_COND DISC_CLASSES SUPP_LHS
## 1 NCKAP5L,234817_at  control     value<cut,value<cut          1,1       18
## 2      MAP7,ATXN8OS  control     value>cut,value<cut          3,1       18
## 3      ZSCAN18,NPR2  control value<cut,cut<value<cut          1,2       19
##   SUPP_RHS ACC_RHS COV_LHS COV_RHS STAB_LHS STAB_RHS   CUT_1   CUT_2
## 1       18 0.97368 0.13740 0.30196        1        1 1.90584 1.64213
## 2       18 1.00000 0.13308 0.29932        1        1 2.51985 2.22742
## 3       19 0.98521 0.14616 0.32895        1        1 2.35647 2.54040
##     CUT_3 CUT_4         PVAL   RISK_PVAL REL_RISK    CONF_INT
## 1     NaN   NaN 4.818175e-06 0.005285147  2.28125 1.273:4.089
## 2     NaN   NaN 4.818175e-06 0.005285147  2.28125 1.273:4.089
## 3 2.59265   NaN 4.818175e-06 0.003949585  2.28125 1.298:4.009
```

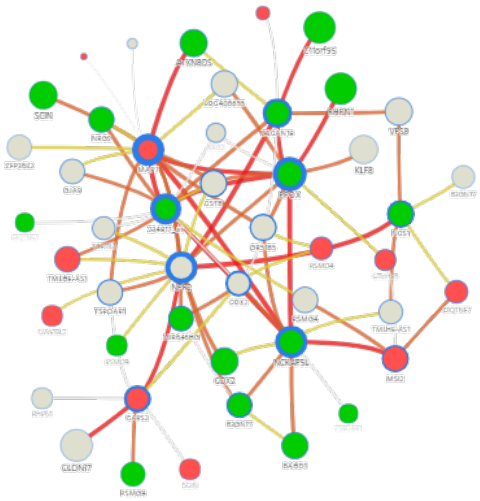

Figure 1: network.

```
table(autconDefault$main[which(autconDefault$main$PVAL <0.05),]$DECISION)
```

```
##
##  autism control
##     108      77
```

e) We obtained 191 rules. Above we can see 3 most significant rules and also we can see the distribution of the classes with more significant rules.

Task 5

At the picture we can see how the network looks like. We found the following strongest connections in the graph MAP7=3 - NCKAP5L=1 (conn: 38.3052), NCKAP5L=1 - PPOX=1 (conn: 38.0476), NCS1=1 - NPR2=2 (conn: 38.0952)

And following most significant nodes: significant nodes: Name: PPOX=1 Edges: 12 Connection: 188.37916 Mean accuracy: 0.941 Mean support: 16.667

Name: MAP7=3 Edges: 16 Connection: 224.4915 Mean accuracy: 0.931 Mean support: 14.875

Name: NPR2=2 Edges: 13 Connection: 193.54575 Mean accuracy: 0.953 Mean support: 15.615

We can find the gene NCKAP5L at the webpage provided. We can see that this is rare single gene mutation and is reported in association with autism. In this article: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5299575/ we can also see that there is a connection between MAP7 gene and autism (and also schizophrenia). According to informations in wikigenes we can see that PPOX gene has some influence on mental health. We didn't find any evidence of strong relation of NPR2 gene to autism.

Appendix

```r
library(R.ROSETTA)
View(autcon)
data("autcon")
table(autcon$decision)
autconDefault = rosetta(autcon)
autconDefault$main
table(autconDefault$main$DECISION)
autconDefault$quality
#CV = 10
#reducer = "Johnson
#discreteMethod = "EqualFrequency"
#discreteParam = 3
#Mean accurancy = 0.821818
autconDefault$main[1:3,]
length(autconDefault$main[which(autconDefault$main$PVAL <0.05),])
saveLineByLine(autconDefault$main, "rules.txt")
save.image(file = "Rosetta_Lab5.RData")
load("Rosetta_Lab5.RData")
```