Assignment 1 (Task 2)

Piotr Filipowicz, Mateusz Kowalewski

10 11 2019

Description

As our clasifier in this task we choosed a random forest classifier. To evaluate it performance we decided to use a confusion matrix. The best model we manage to achieve use 466 features, so we were able do decrease number of features to half. This model achieve accuracy of 92.8%, sensitivity (Proper clasifying of DLBCL) of 90.9% and specificity (Proper clasifying of FL) of 100%. Code:

```
DLBCL = read.csv("DLBCL.csv")

inTraining <- createDataPartition(DLBCL$class, p = .80, list = FALSE)
training <- DLBCL[ inTraining,]
testing  <- DLBCL[-inTraining,]

randomforest <- train(training[2:1072],training$class,
                  method="rf", metric="ROC",
                  preProcess = c("center", "scale"),
                  tuneLength = 10,
                  trControl = trainControl(method = "cv",number = 3,
                  classProbs=TRUE, summaryFunction=twoClassSummary))
```

Discussion

1). The feature selection was succesfull. We manage to decrease number of features to 466.

2). We may try to implement diffrent feature selection algorithms, for example: Genetic Algorithm Feature Selection, Feature Selection using Simulated Annealing, Recursive Feature Elimination.

3). We may extend training data, but we would have to decrease test data or add more entries. Also we can try to use cross-validation. In this task we used 3-cross-validation. Moreover we could add “early stopping point” to stop when classifier is satisfying for our problem

4). Total number of features, number of features after selection, accuracy of classifier, specificity and sensitivity

5). After comparing, we can observe that a lot of features are repeated in both lists. Moreover features with higher correlation with target variable are more likely to appear in the list of final set of features.

```r
#Code for question 5

x <- c(2:1072)
max_index <- c()
for(val in x){
  max_index[val] <- biserial.cor(DLBCL[,val], DLBCL$class)
}
y <- order(max_index, decreasing=TRUE)[1:466]
#DLBCL[0,y]
```