# Homework0411

## 纪静远 SA19234098

### 数据导入与清洗

```
library("R.matlab")
library(caret)
library(randomForest)
myData=data.frame(readMat('/home/crick/R/Data/dataSet.mat'))
```

数据集为斑马鱼幼鱼运动学参数与运动分类信息

```
good=complete.cases(myData)
myData=myData[good,]

myLabel=myData$dataSet.109
myFeature=myData[,-109]
```

去除无效变量，分类保存为Label与Feature，
其中Faeture含有108个变量

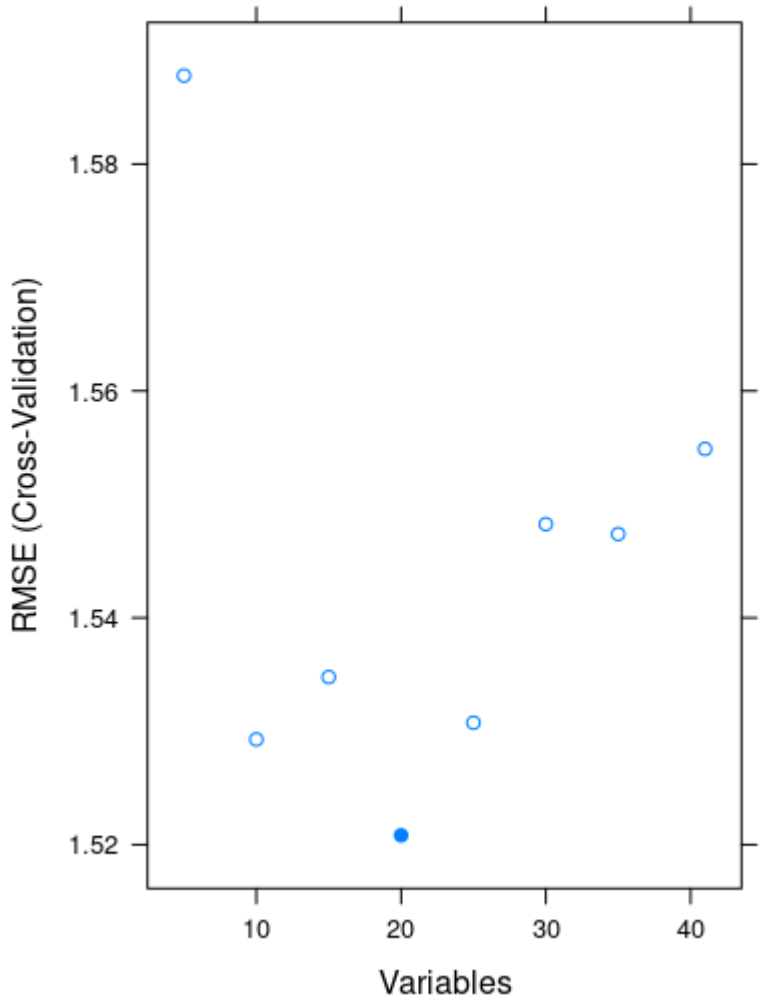
```
descrCorr = cor(myFeature)
highCorr = findCorrelation(descrCorr, 0.90)
myFeature=myFeature[,-highCorr]
comboInfo = findLinearCombos(myFeature)
myFeature=myFeature[, -comboInfo$remove]
```

去除高度相关变量与多重线性相关变量，
Feature剩余41个变量

```
subsets = c(5,10,15,20,25,30,35,41)

ctrl= rfeControl(functions = rfFuncs, method = "cv",verbose = FALSE, returnResamp = "final")
Profile = rfe(myFeature, myLabel, sizes = subsets, rfeControl = ctrl)
print(Profile)
plot(Profile)
```

随机森林建模并计算特征效能




故选择20个自变量

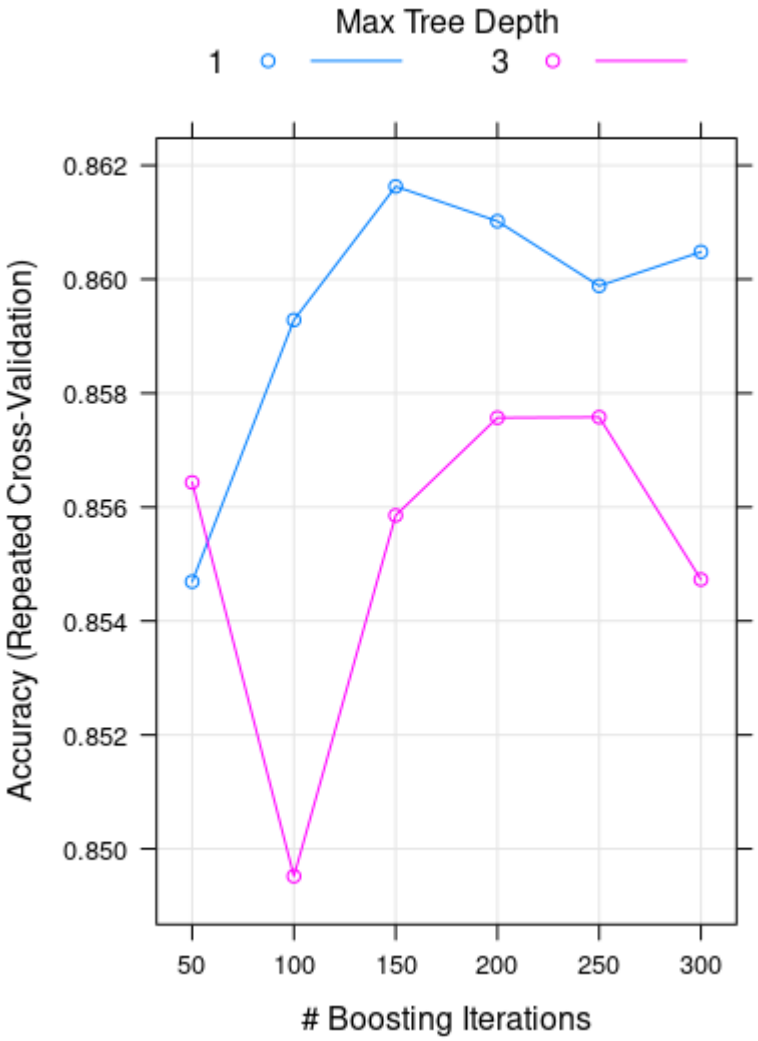
```
myFeature=myFeature[,Profile$optVariables]

inTrain = createDataPartition(myLabel, p = 3/4, list = FALSE)
trainFeature=as.data.frame(myFeature[inTrain,])
testFeature=as.data.frame(myFeature[-inTrain,])
trainLabel=as.character(myLabel[inTrain])
testLabel=as.character(myLabel[-inTrain])
testLabel=factor(testLabel)
```

划分训练集和测试集

```
fitControl = trainControl(method = "repeatedcv", number = 5, repeats = 5,returnResamp = "all")
gbmGrid = expand.grid(.interaction.depth = c(1, 3),.n.trees = c(50, 100, 150, 200, 250, 300),.shrinkage = 0.1
gbmFit1 = train(trainFeature,trainLabel,method = "gbm",trControl = fitControl,tuneGrid = gbmGrid,verbose = FA
```

训练提升树分类模型




```
pre=predict(gbmFit1, newdata = testFeature)
confusionMatrix(testLabel,pre)
```

预测以及计算混淆矩阵

Confusion Matrix and Statistics

Overall Statistics

```
               Accuracy : 0.8435
                 95% CI : (0.764, 0.9045)
    No Information Rate : 0.6435
    P-Value [Acc > NIR] : 1.624e-06
                  Kappa : 0.719
```

Mcnemar's Test P-Value : NA

test