2017年《数据分析工具实践》

随机森林的R实现之randomForest包

二学位　许昕　1600022716

指导老师　孙惠平

北京大学

PEKING UNIVERSITY

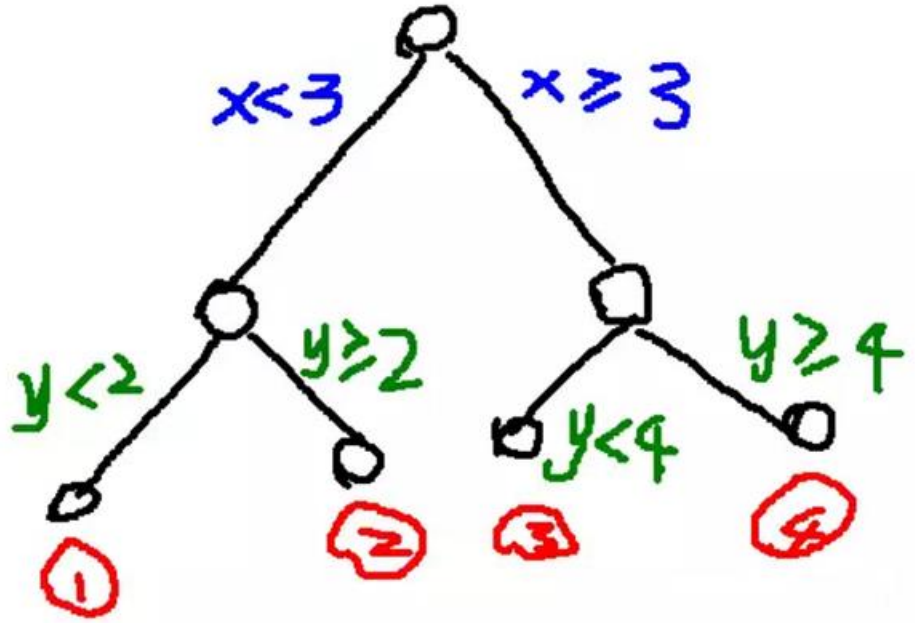
1 什么是随机森林

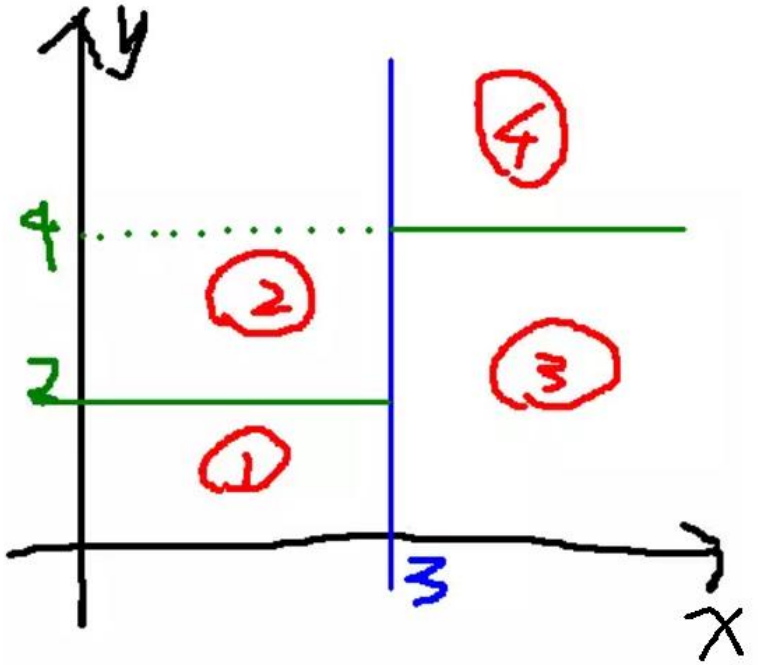
1.1 什么是随机森林

- 顾名思义，随机森林就是用随机的方式建立一片森林，森林由很多的决策树组成，随机森林的每一棵决策树之间是没有关联的
- 在得到森林之后，当有一个新的输入样本进入的时候，就让森林中的每一棵决策树分别进行一下判断，看看这个样本应该属于哪一类（对于分类算法），然后看看哪一类被选择最多，就预测这个样本为那一类
- 随机森林既可以处理属性为离散值的量，也可以处理属性为连续值的量。另外，随机森林还可以用来进行无监督学习聚类和异常点检测

1.2 什么是随机森林

- 随机森林由决策树组成，决策树实际上是将空间用超平面进行划分的一种方法，每次分割的时候，都将当前的空间一分为二





1.3 什么是随机森林

- 随机森林是一种利用多棵决策树对数据进行判别与分类的方法，它在对数据进行分类的同时，还可以给出各个变量的重要性评分，评估各个变量在分类中所起的作用
- 可以这样比喻随机森林算法：每一棵决策树就是一个精通某一个窄领域的专家，这样在随机森林中就有了很多个精通不同领域的专家，对一个新的问题（新的输入数据），可以用不同的角度去看待它，最终由各个专家投票得到结果（群体智慧）

2

随机森林的应用

2.1 随机森林的应用

- randomForest函数的基本书写格式为：

  randomForest(输出变量名~输入变量名, data=数据框名, mtry=k, ntree=M, importance=TRUE)

- 数据事先组织在data参数指定的数据框中
- 参数mtry用于指定决策树各节点的输入变量个数k
- 参数ntree用于指定随机森林包含M棵决策树，默认为500
- 参数importance=TRUE表示计算输入变量对输出变量重要性的测度值

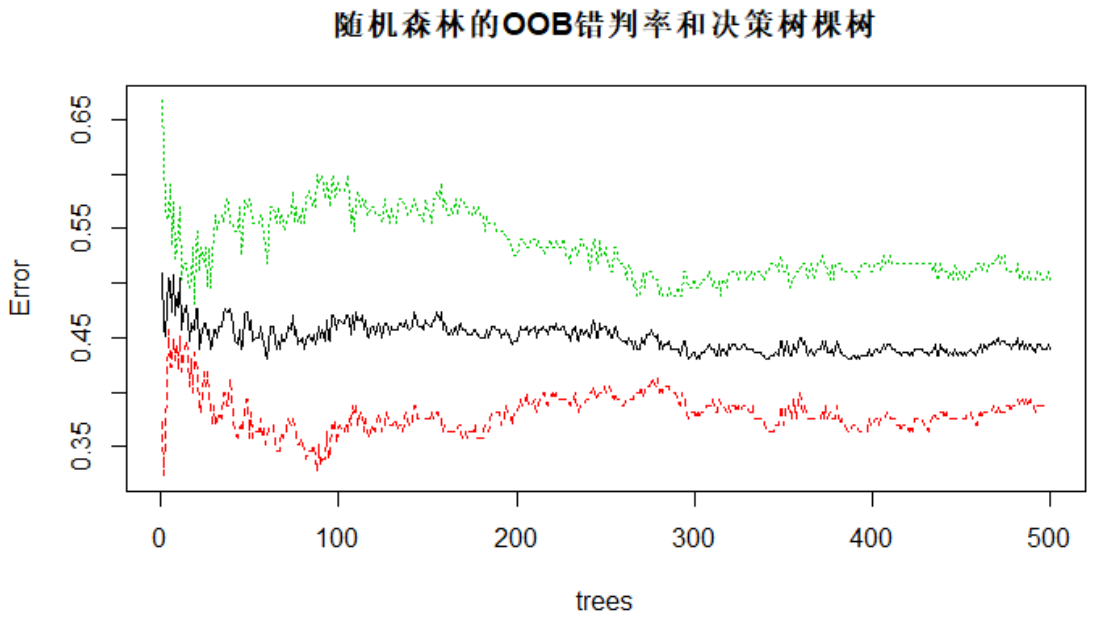
2.2 随机森林的应用

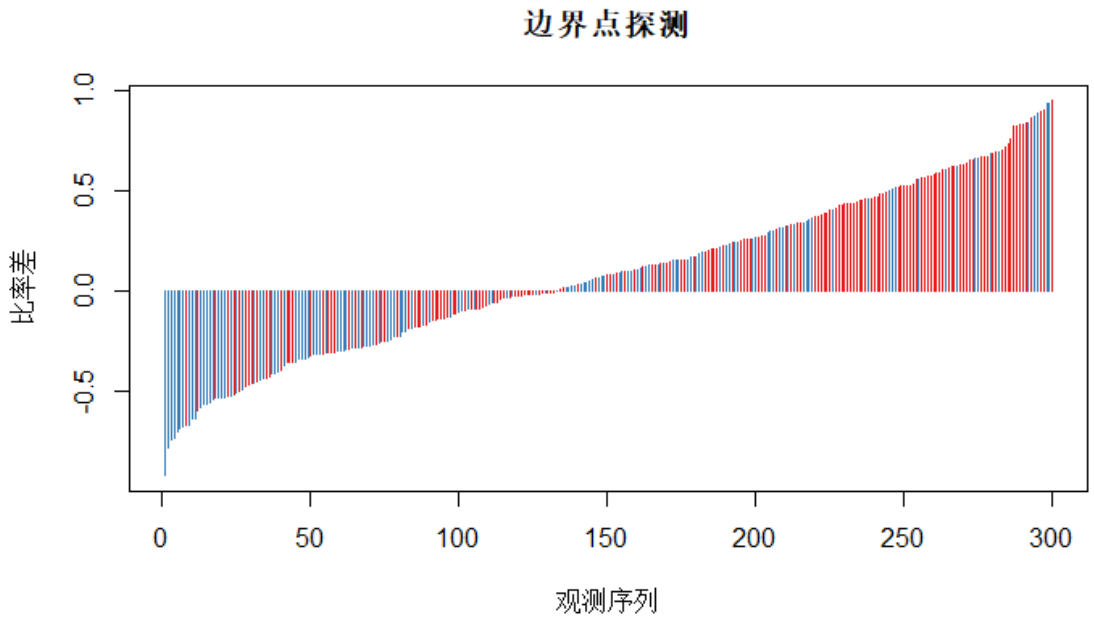
- 随机森林共建立了500棵决策树，每个节点的候选输入变量个数为2。基于袋外观测OOB的预测错判率为42.67%
- 以第1个观测为例：有63%的决策树投票给NO类，37%投票给YES类。它有189次作为OOB未进入训练样本集

```
> MailShot<-read.table(file="MailShot.txt",header=TRUE)
> MailShot<-MailShot[,-1]
> set.seed(12345)
> rFM<-randomForest(MAILSHOT~.,data=MailShot,importance=TRUE)
> rFM

Call:
 randomForest(formula = MAILSHOT ~ ., data = MailShot, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 44%
Confusion matrix:
     NO YES class.error
NO  101  64   0.3878788
YES  68  67   0.5037037
> head(rFM$votes)
         NO       YES
1 0.5459459 0.4540541
2 0.3440367 0.6559633
3 0.8324324 0.1675676
4 0.8032787 0.1967213
5 0.3602151 0.6397849
6 0.3791209 0.6208791
> head(rFM$oob.times)
[1] 185 218 185 183 186 182
```

2.3 随机森林的应用

```
> plot(rFM,main="随机森林的OOB错判率和决策树棵树")
> plot(margin(rFM),type="h",main="边界点探测",xlab="观测序列",ylab="比率差")
```





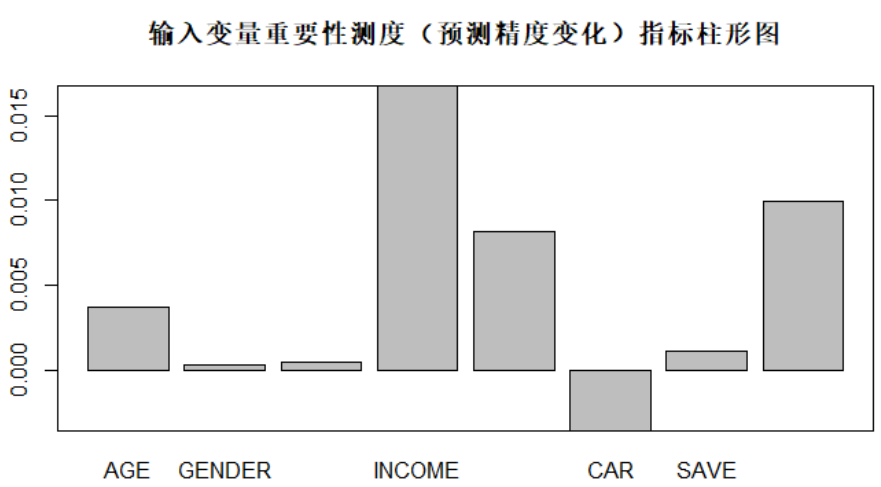
2.4 随机森林的应用

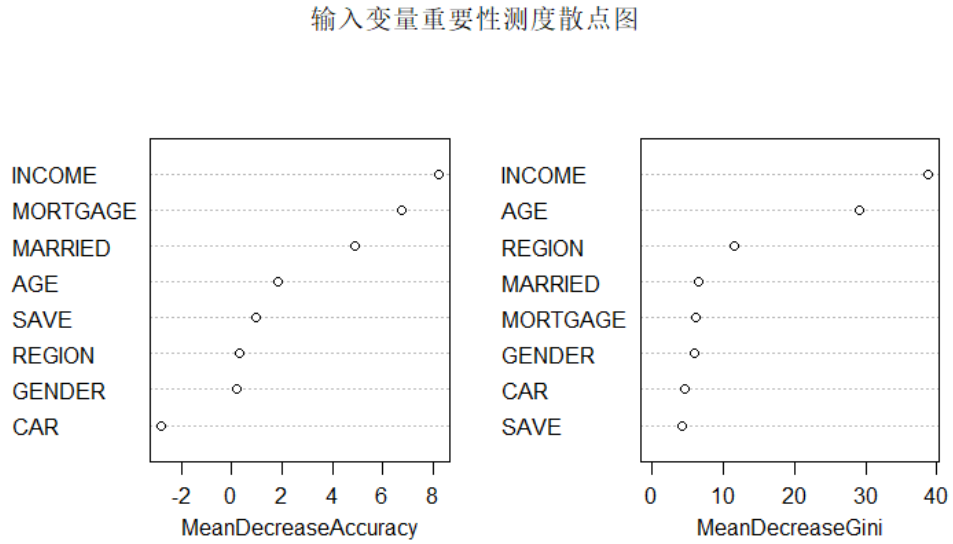
- 可利用treesize函数显示随机森林中各决策树的大小
- 可利用getTree函数抽取随机森林中的某棵树并浏览其结构

```
> head(treesize(rFM))
[1] 65 76 49 75 48 68
> head(getTree(rfobj=rFM,k=1,labelVar=TRUE))
  left daughter right daughter split var split point status prediction
1             2              3       AGE        47.5      1       <NA>
2             4              5   MARRIED         1.0      1       <NA>
3             6              7    REGION         7.0      1       <NA>
4             8              9    REGION         1.0      1       <NA>
5            10             11    REGION         4.0      1       <NA>
6            12             13    INCOME     42903.5      1       <NA>
```

2.5 随机森林的应用

```
> barplot(rFM$importance[,3],main="输入变量重要性测度（预测精度变化）指标柱形图")
> box()
> importance(rFM,type=1)
          MeanDecreaseAccuracy
AGE                  1.8484166
GENDER               0.2103645
REGION               0.3246826
INCOME               8.2267037
MARRIED              4.8779590
CAR                 -2.7974752
SAVE                 0.9860112
MORTGAGE             6.7435090
> varImpPlot(x=rFM,sort=TRUE,n.var=nrow(rFM$importance),main="输入变量重要性测度散点图")
```





3 随机森林练习题

3.1 随机森林练习题

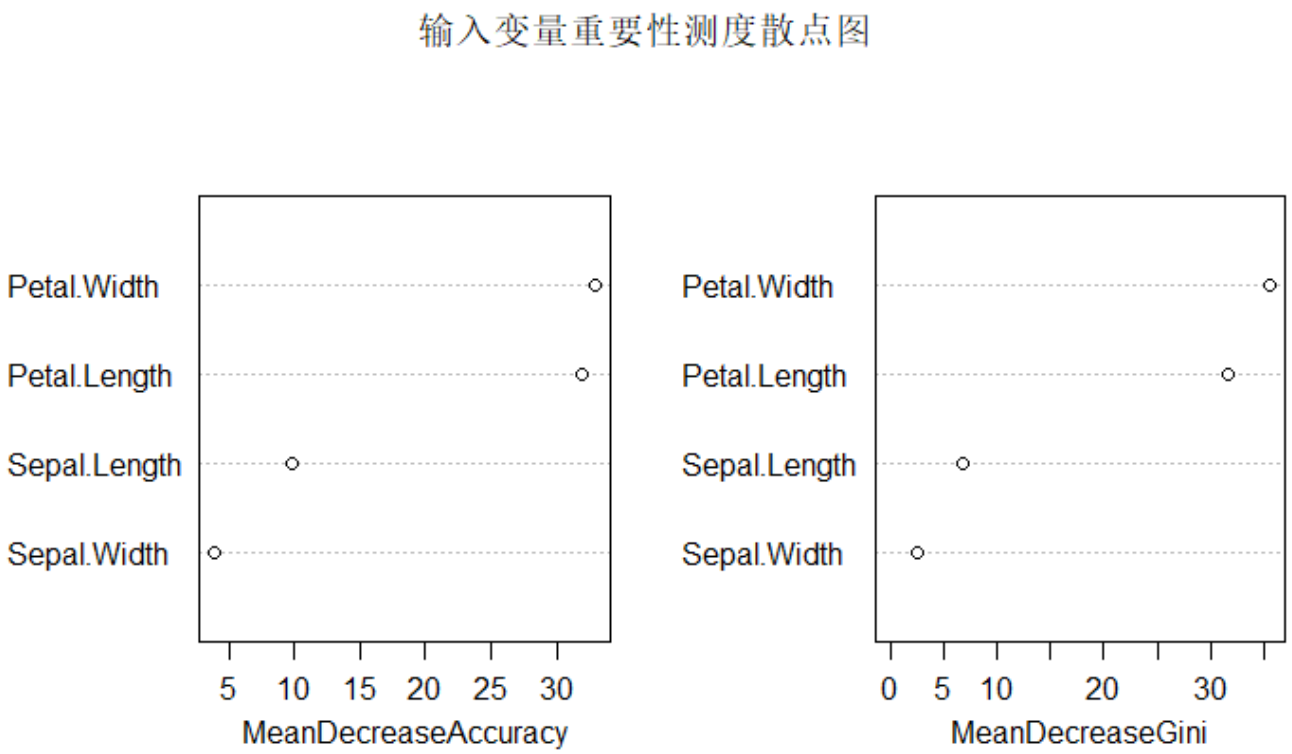
一、针对iris数据集，通过随机森林的算法，根据一些特征（花瓣和花萼的长、宽）来预测植株的种类。

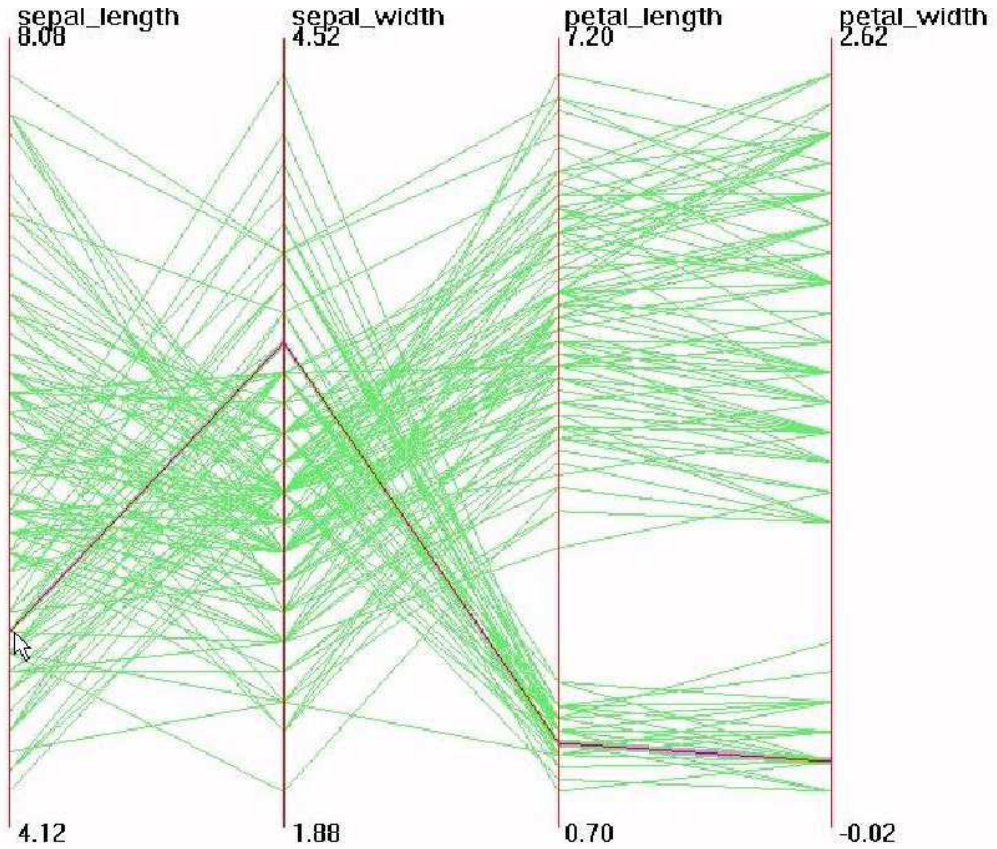
二、针对cancer数据集，通过随机森林的算法，根据一些特征（肿块厚度、细胞大小的均匀性等）来预测癌症的种类（良性 or 恶性）。

3.2 练习题一解答（1）

```
> data("iris")
> set.seed(12345)
> ind<-sample(2,nrow(iris),replace=TRUE,prob=c(0.8,0.2))
> (iris.rf<-randomForest(Species~.,iris[ind==1,],importance=TRUE))

Call:
 randomForest(formula = Species ~ ., data = iris[ind == 1, ],      importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 6.03%
Confusion matrix:
           setosa versicolor virginica class.error
setosa         42          0         0  0.00000000
versicolor      0         35         3  0.07894737
virginica       0          4        32  0.11111111
> iris.pred<-predict(iris.rf,iris[ind==2,])
> table(observed=iris[ind==2,"Species"],predicted=iris.pred)
            predicted
observed     setosa versicolor virginica
  setosa          8          0         0
  versicolor      0         12         0
  virginica       0          1        13
> varImpPlot(x=iris.rf,sort=TRUE,n.var=nrow(iris.rf$importance),main="输入变量重要性测度散点图")
```

3.2 练习题一解答（2）


输入变量重要性测度散点图
Petal.Width
Petal.Length
Sepal.Length
Sepal.Width
5 10 15 20 25 30
MeanDecreaseAccuracy
Petal.Width
Petal.Length
Sepal.Length
Sepal.Width
0 5 10 20 30
MeanDecreaseGini

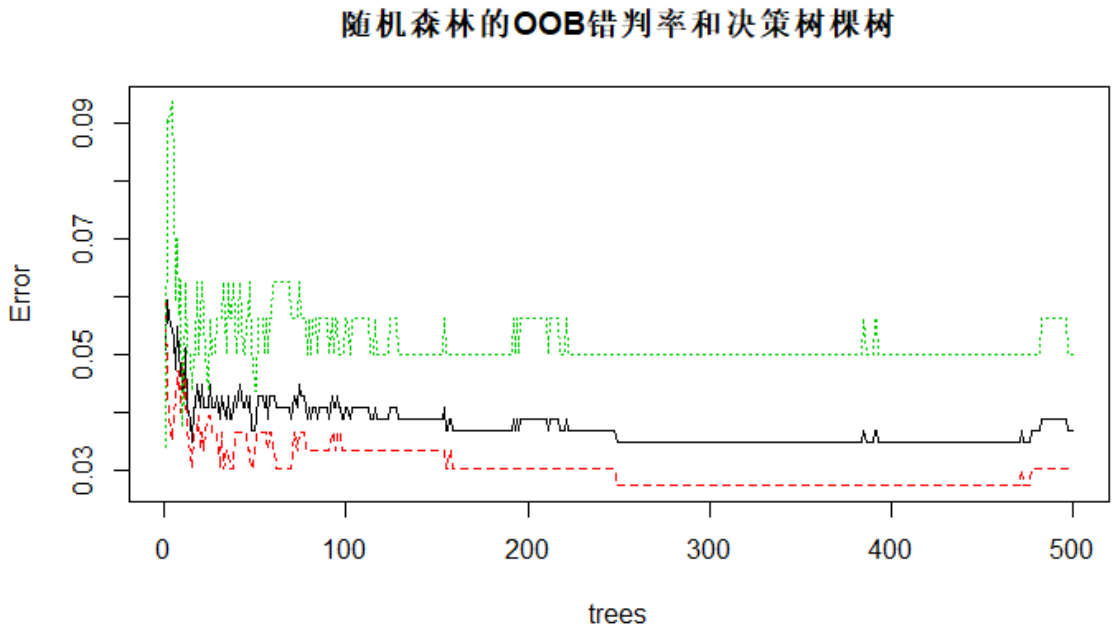

sepal_length
8.08
4.12
sepal_width
4.52
1.88
petal_length
7.20
0.70
petal_width
2.62
-0.02

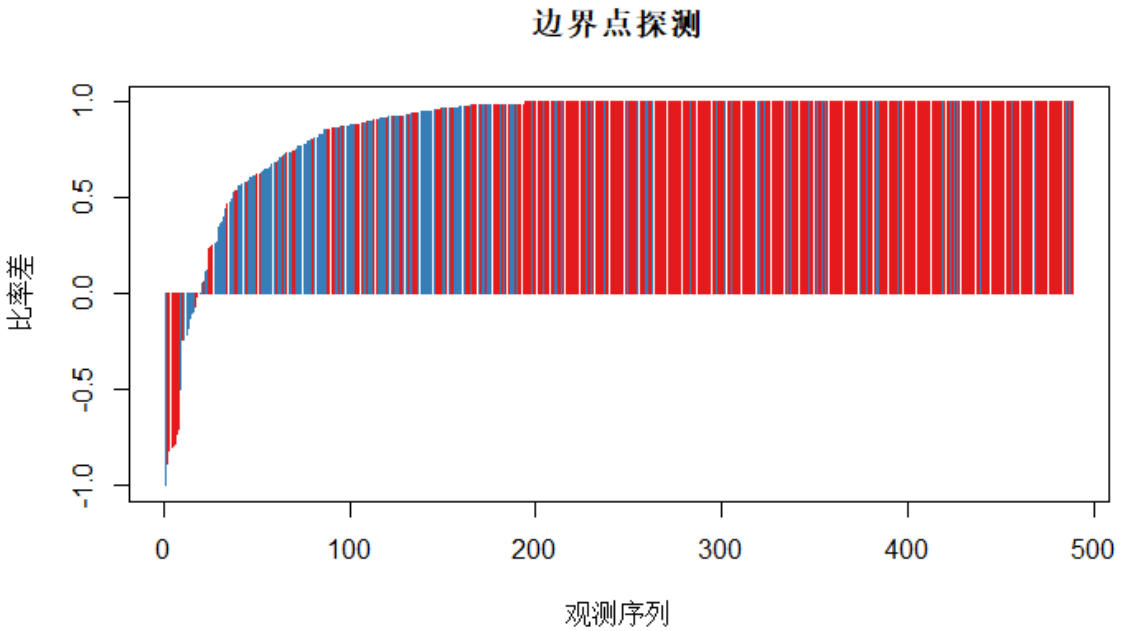
3.3 练习题二解答（1）

```
> breast <- read.table(file="cancer.txt", sep=",", header=FALSE, na.strings="?")
> names(breast)<-c("ID","clumpThickness","sizeUniformity","shapeUniformity","maginalAdhesion","singl
eEpithelialCellSize","bareNuclei","blandChromatin","normalNucleoli", "mitosis", "class")
> df <- breast[-1]
> df$class <- factor(df$class, levels=c(2,4),labels=c("benign", "malignant"))
> set.seed(1234)
> train <- sample(nrow(df), 0.7*nrow(df))
> df.train <- df[train,]
> df.validate <- df[-train,]
> table(df.train$class)

   benign malignant
      329       160
> table(df.validate$class)

   benign malignant
      129        81
> set.seed(1234)
> (fit.forest<-randomForest(class~.,data=df.train,na.action=na.roughfix,importance=TRUE))

Call:
 randomForest(formula = class ~ ., data = df.train, importance = TRUE,      na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 3.68%
Confusion matrix:
          benign malignant class.error
benign       319        10  0.03039514
malignant      8       152  0.05000000
```

3.3 练习题二解答（2）

```
> plot(fit.forest,main="随机森林的OOB错判率和决策树棵树")
> plot(margin(fit.forest),type="h",main="边界点探测",xlab="观测序列",ylab="比率差")
> forest.pred <- predict(fit.forest, df.validate)
> (forest.perf <- table(df.validate$class, forest.pred,dnn=c("Actual", "Predicted")))
           Predicted
Actual      benign malignant
  benign       117         3
  malignant      1        79
```





Thank you!

北京大學
PEKING UNIVERSITY