HW1

0513250 邱郁雯

1. 切割訓練資料與測試資料，比例為 80%：20%。

切割 iris 資料庫，80%為訓練資料(iris.train)，20%為測試資料(iris.test)。

```
n<- 0.2*nrow(iris)
index <- sample(1:nrow(iris),n)
iris.train <- iris[-index,]
iris.test <- iris[index,]
```

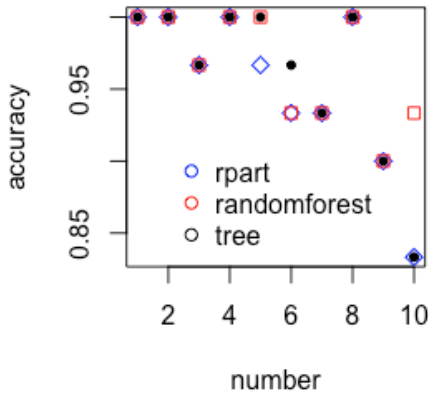
圖一

2. 分別利用 tree 套件（不修剪）、rpart 套件（不修剪）、與 randomForest 進行 10 次測試，並比較其準確度（包含準確度的變化）。

將三個套件在迴圈進行 10 次測試後，tree、rpart、randomforest 的準確度為圖二，並將測試次數與準確度畫出散佈圖（圖三），可看出測試次數越多，準確度沒有明顯提升，在一開始測試時準確度就已經很高，所以準確度變化不大。由圖四可知，tree 平均準確度為 0.96，rpart 平均準確度為 0.9533，randomforest 平均準確度為 0.9667，所以這三個套件的準確度 randomforest > tree > rpart。

```
> df
   accuracy.atr accuracy.arp accuracy.arf
1     1.0000000    1.0000000    1.0000000
2     1.0000000    1.0000000    1.0000000
3     0.9666667    0.9666667    0.9666667
4     1.0000000    1.0000000    1.0000000
5     1.0000000    0.9666667    1.0000000
6     0.9666667    0.9333333    0.9333333
7     0.9333333    0.9333333    0.9333333
8     1.0000000    1.0000000    1.0000000
9     0.9000000    0.9000000    0.9000000
10    0.8333333    0.8333333    0.9333333
```

圖二



圖三

```
> summary(df)
  accuracy.atr      accuracy.arp      accuracy.arf
 Min.   :0.8333   Min.   :0.8333   Min.   :0.9000
 1st Qu.:0.9417   1st Qu.:0.9333   1st Qu.:0.9333
 Median :0.9833   Median :0.9667   Median :0.9833
 Mean   :0.9600   Mean   :0.9533   Mean   :0.9667
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
```

圖四

```
for(i in 1:10){
  n<- 0.2*nrow(iris)
  index <- sample(1:nrow(iris),n)
  iris.train <- iris[-index,]
  iris.test <- iris[index,]
  #tree
  iris.tree1 <- tree(Species ~ .,data=iris.train)
  test.tr<- predict(iris.tree1,iris.test, type = "class")
  compare.tr <- ifelse(test.tr==iris.test$Species,1,0)
  x<-sum(compare.tr)/length(compare.tr)
  accuracy.atr[i]<-x
  #rpart
  iris.tree2 <- rpart(Species~., data = iris.train, cp=0)
  test.rp<- predict(iris.tree2,iris.test, type = "class")
  compare.rp <- ifelse(test.rp==iris.test$Species,1,0)
  y<-sum(compare.rp)/length(compare.rp)
  accuracy.arp[i]<-y
  #randomforest
  iris.rf<-randomForest(Species~.,data = iris.train)
  test.rf<- predict(iris.rf,iris.test, type = "class")
  compare.rf <- ifelse(test.rf==iris.test$Species,1,0)
  z<-sum(compare.rf)/length(compare.rf)
  accuracy.arf[i]<-z
}
```

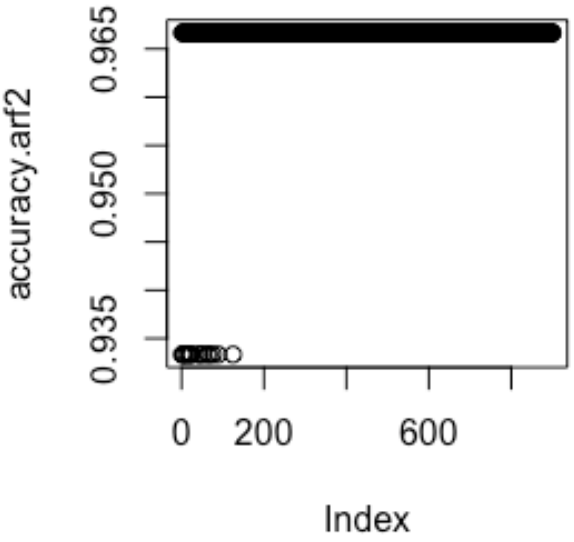
圖五

3. 利用 randomForest 套件，比較不同森林大小的準確度，並建議最合適的森林大小（分類樹的個數）。

使用 for 迴圈測試最佳的分類樹個數，從 1 顆到 900 顆測試，並且將每次測試之準確度畫成散佈圖，結果如圖七，1 到 900 顆的分類樹準確度皆在 90% 以上，並且當分類數在 200 顆後，準確率高達 100%。根據結果，我認為最佳分類樹的個數大於 200 都很適合。

```r
nrf<- 0.2*nrow(iris)
index.rf <- sample(1:nrow(iris),nrf)
iris.trainrf <- iris[-index.rf,]
iris.testrf <- iris[index.rf,]
accuracy.arf2<-c(1:800)
for (i in 1:900) {
  iris.rf2<-randomForest(Species~.,data = iris.trainrf,ntree=i)
  test.rf2<- predict(iris.rf2,iris.testrf, type = "class")
  compare.rf2 <- ifelse(test.rf2==iris.testrf$Species,1,0)
  w<-sum(compare.rf2)/length(compare.rf2)
  accuracy.arf2[i]<-w
}
head(accuracy.arf2)
plot(accuracy.arf2)
```

圖六



圖七