**20210512**

作業1：請針對 Dose 利用 G1 進行線性迴歸,並進行兩倍交叉驗證(Two-fold cross-validation)。 我們假設回歸模型預測出來的 Dose 與實際數據的 Dose 相差在0.02 以內時為正確預測, 請計算預測率?

最終預測率為0.45



Code:

## hw1

# 請讀入partA\_cv.csv

partA\_cv <- read.csv(file.choose())

#

random\_index <- sort(sample(100, size=50, replace=F))

A <- partA\_cv[random\_index, ]

B <- partA\_cv[- random\_index, ]

# A=train B=test

train <- A

test <- B

fit <- lm(Dose~G1, data=train)

pre\_fit <- predict(fit, test)

accuracy <- c(abs(pre\_fit - test$Dose) < 0.02)

# A=test B=train

train <- B

test <- A

fit <- lm(Dose~G1, data=train)

pre\_fit <- predict(fit, test)

accuracy <- c(accuracy, abs(pre\_fit - test$Dose) < 0.02)

# calculate probability of accuracy

prob\_accuracy <- sum(accuracy) / length(accuracy)

prob\_accuracy

作業2：我們進行了一次微陣列實驗,假設人類擁有 20000 個基因(編碼為1:20000),其中 Geneset A 為 Column A 所標記,在我們找到的 100 個 Identified

genes 中 Geneset A 佔了 4 個

第一小題:請計算找到的基因在 GenesetA 的超幾何機率?



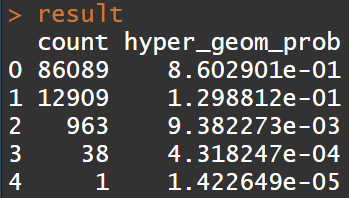
第二小題:請透過排列法隨機抽取 100 個基因 100000 次並測試 Geneset A 顯著(p<0.05)的次數是?與這個次數比較後,你認為 Geneset A 與實驗是否有顯著相

關?

第一行：Geneset A 佔的Identified genes 數

第二行：次數

第三行：超幾何機率





透過排列法隨機抽取100個基因100000次並測試Geneset A顯著(p<0.05)的次數有1002次達到顯著，原先實驗Geneset A佔的Identified genes (4個) 遠大於重複100000次，依照排列法結果p-value < 10−5，因此我認為 Geneset A 與實驗有顯著相關。

Code:

## hw2

choose(30, 4) \* choose(19970, 96) / choose(20000, 100)

# import data

partB\_cv <- read.csv(file.choose())

# data cleaning

geneset\_A <- partB\_cv$Geneset.A[!is.na(partB\_cv$Geneset.A)]

all\_gene <- partB\_cv$Identified.genes

# select

match\_gene <- c()

for(i in 1:100000){

samples <- sample(c(1:20000), size = 100)

match\_gene <- c(match\_gene, length(intersect(samples, geneset\_A)))

}

count <- table(match\_gene)

# 計算超幾何機率

prob\_hypergeom <- sapply(c(0:(length(count) - 1)), function(x) choose(30, x) \* choose(19970, 100 - x) / choose(20000, 100))

result <- cbind(count, hyper\_geom\_prob)

result

# p < 0.05

sum(count[3:length(count)])

**20210519**

作業1：請讀入 Data\_file.csv,檔案第一行是樣本的型態 (Cancer, Normal),且檔案內共有 54675 個探針,請自由發揮篩選方法選出你認為最有用的探針並透過SVM 和 CART 建立預測器,最後請透過兩倍交互驗證測試預測器的準確度

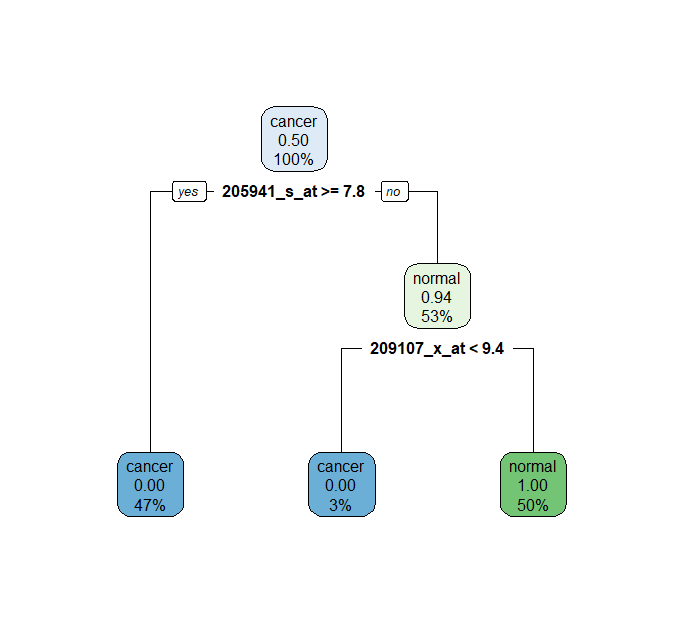
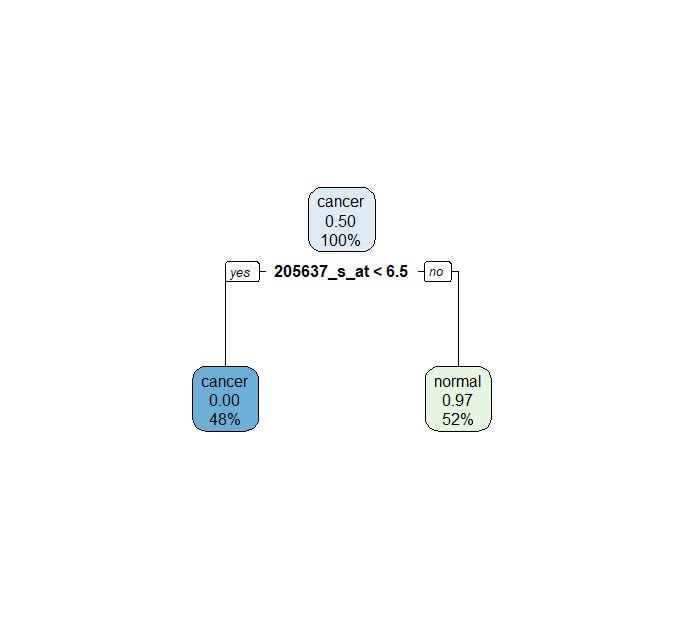
利用t-test以p-value < 10^(-15) 篩選了906個探針

SVM的準確度為0.967

CART的準確度為0.925

此為CART的兩倍交互驗證的兩模型

Code:

## hw3

library(e1071)

library(rpart)

install.packages("rpart.plot")

library(rpart.plot)

### import data\_file

data <- read.csv(file.choose())

# 取出probe名字

probename <- data$Type

# 轉置

data <- as.data.frame(t(data[, - 1]))

names(data) <- probename

phenotype <- as.factor(rep(c("cancer","normal"), each=60))

data <- cbind(phenotype, data)

### 挑probe

# 儲存顯著的p-value和位點 (注意這邊的是於data的行數)

probe <- c()

for(i in 2:54676){

p <- t.test(data[1:60, i], data[61:120, i])$p.value

probe <- rbind(probe, c(i, p))

}

probe <- as.data.frame(probe)

# 大約挑個900個p較小的probe (注意這邊的是於data的行數)

# 這裡隨意改

probe\_select <- subset(probe, probe$V2 < 10 ^ (- 15))

probe\_input <- probe\_select$V1

# 把data只取想要的probe 之後才好放入model

data\_pro <- data[, c(1, probe\_input)]

### cross validation分2組

C <- sample(c(1:60), size=30)

N <- sample(c(61:120), size=30)

data\_A <- data\_pro[c(C, N), ]

data\_B <- data\_pro[- c(C, N), ]

### SVM

# 1

SVM\_model1 <- svm(phenotype ~ ., data = data\_A, kernel = "linear")

SVM\_output1 <- predict(SVM\_model1, data\_B)

SVM\_result1 <- table(SVM\_output1, data\_B$phenotype, dnn = c("predict", "true"))

# 2

SVM\_model2 <- svm(phenotype ~ ., data = data\_B, kernel = "linear")

SVM\_output2 <- predict(SVM\_model2, data\_A)

SVM\_result2 <- table(SVM\_output2, data\_A$phenotype, dnn = c("predict", "true"))

# result & acc

SVM\_result <- SVM\_result1 + SVM\_result2

SVM\_accuracy <- sum(diag(SVM\_result)) / sum(SVM\_result)

SVM\_accuracy

### CART

# 1

CART\_model1 <- rpart(phenotype ~ ., data = data\_A, minsplit = 5)

CART\_output1 <- predict(CART\_model1, data\_B, type = "class")

CART\_result1 <- table(CART\_output1, data\_B$phenotype, dnn = c("predict", "true"))

rpart.plot(CART\_model1)

# 2

CART\_model2 <- rpart(phenotype ~ ., data = data\_B, minsplit = 5)

CART\_output2 <- predict(CART\_model2, data\_A, type = "class")

CART\_result2 <- table(CART\_output2, data\_A$phenotype, dnn = c("predict", "true"))

rpart.plot(CART\_model2)

# result & acc

CART\_result <- CART\_result1 + CART\_result2

CART\_accuracy <- sum(diag(CART\_result)) / sum(CART\_result)

CART\_accuracy