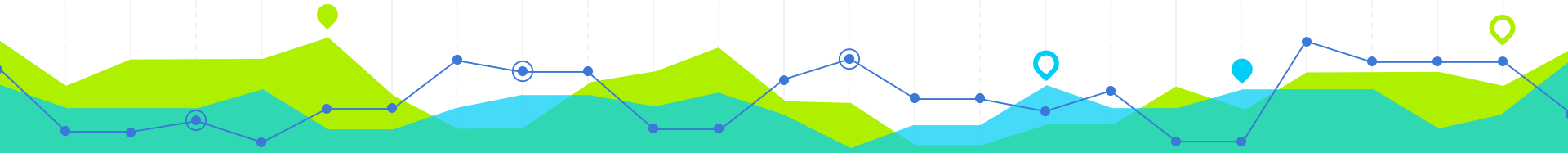


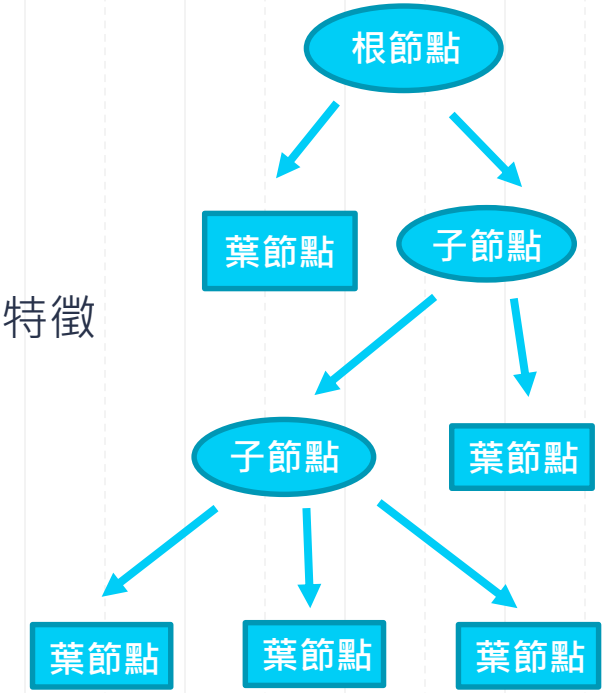
# Random Forest

7109018032謝綺珊
7109018037李蕙君

# Decision Tree (決策樹)

- 模仿人類做決策的過程
- 分為root和sub-trees兩部分
- 透過信息增益(information gain)來判斷用於分割的特徵
  例如 : Gini impurity、entropy

$$IG(D_p, \mathrm{f}) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$



# Decision Tree (決策樹)

- Gini impurity : $p(i|t)$代表在節點t，屬於類別c的比例

$$I_G(t) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

- Entropy :對不確定性的測量

$$I_H(t) = -\sum_{i=1}^{c} p(i|t) \log_2 p(i|t)$$

- Entropy比不純度更加敏感，且Entropy的計算比不純度要緩慢，但是在實際使用中， Entropy和不純度的效果基本上是相同的。

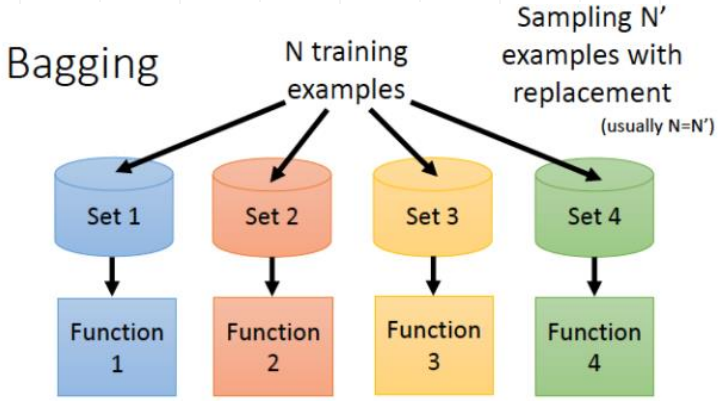
# Decision Tree (決策樹)

- 優點 : 1.樹會自動選擇在模型中用於決定分類的變數
  2.模型直觀，便於理解，應用廣泛
  3.可以處理混合特徵和具遺失的資料
  4.訓練和預測時，效率較高
  5.小型的樹很容易去解釋

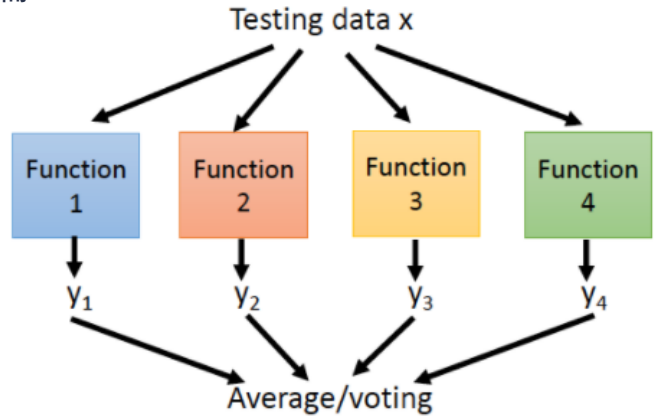
- 缺點 : 1.缺少足夠的理論支持
  2.樹通常預測的不太好
  3.大型的樹很難去解釋

# Bagging (Bootstrap Aggregation)

- 從 N 筆 training data 中，做 sampling 組成 M 個 dataset 每個 dataset 裡面有 N' 筆資料，使用 sampling 的方法建出很多資料集，訓練出多個 function




- 接著再把我們訓練出來的 function 跑出來得結果拿出來整合，得到最後的結論




- 對於複雜或容易過擬的模型，bagging是有幫助的

# OOB (Out-of-Bag) Error Estimate

- 使用bootstrap 生成訓練資料，因此原始訓練資料中有一部份不會出現在訓練資料中，這些資料便稱為$OOB$資料
- 對每一棵樹都透過$OOB$資料去估計誤差，將森林中每一棵樹的$OOB$誤差取平均，便得到隨機森林的$OOB$誤差估計

$$\hat{r}_{\mathrm{rf}}^{(i)}(x_i) = \frac{1}{B_i} \sum_{b\,:\,w_{b\,i}^{*}=0} \hat{r}_b(x_i) \qquad \mathrm{err}_{\mathrm{OOB}} = \frac{1}{n} \sum_{i=1}^{n} L[y_i, \hat{r}_{\mathrm{rf}}^{(i)}(x_i)]$$

| train | $f_1$ | $f_2$ | $f_3$ | $f_4$ |
|---|---|---|---|---|
| $x^1$ | O | X | O | X |
| $x^2$ | O | X | X | O |
| $x^3$ | X | O | O | X |
| $x^4$ | X | O | X | O |

Out-of-bag validation for bagging

- Using RF = $f_2$+$f_4$ to test $x^1$
- Using RF = $f_2$+$f_3$ to test $x^2$
- Using RF = $f_1$+$f_4$ to test $x^3$
- Using RF = $f_1$+$f_3$ to test $x^4$

Out-of-bag (OOB) error

Good error estimation of testing set

# Random Forest (隨機森林)

- Random Forest = Bagging + Decision Tree
- 一種集成學習(ensemble learning)技術
- 結合多棵樹，且加入隨機分配的訓練資料，以大幅增進最終的運算結果，並且在節點則隨機選擇特徵子集來分割資料，以此降低隨機森林的錯誤率
- 產生多顆具差異性的樹可利用bagging或boosting
- 每次分類，只選擇$p$個變數中的$m$個隨機變數子集合。通常選擇$m = \sqrt{p}$ ($R$ default), $or$ $m = p/3$.

# Random Forest (隨機森林)

- Algorithm :

1.給定訓練資料 $d$=$(X, y)$ ，決定分類變數個數$m \leq p$和樹的個數$B$

2.對每一棵樹 $b = 1, \ldots, B$:

(a)在$n$列訓練資料中隨機重複抽出$n$列當次訓練資料 $d_b^*$

(b)在每個分類之前隨機採樣$m$個變數，使用$d_b^*$去生成樹 $\hat{r}_b(x)$

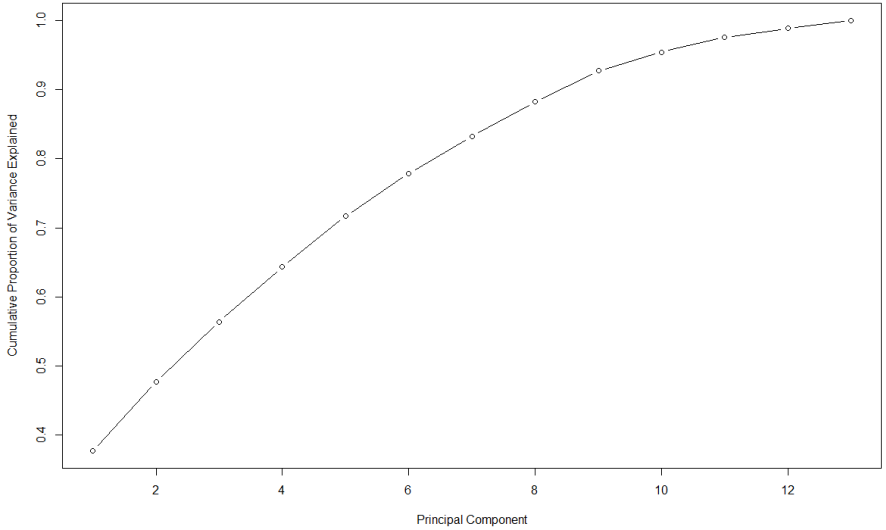
3.將預測資料代入隨機森林模型

$$\hat{r}_{rf}(x_0) = \frac{1}{B}\sum_{b=1}^{B}\hat{r}_b(x_0)$$

4.計算訓練資料中對沒有被bootstrap抽樣到的資料$i$的反應變數 $y_i$的OO$B_i$誤差，整體$OOB$誤差會是OO$B_i$的平均

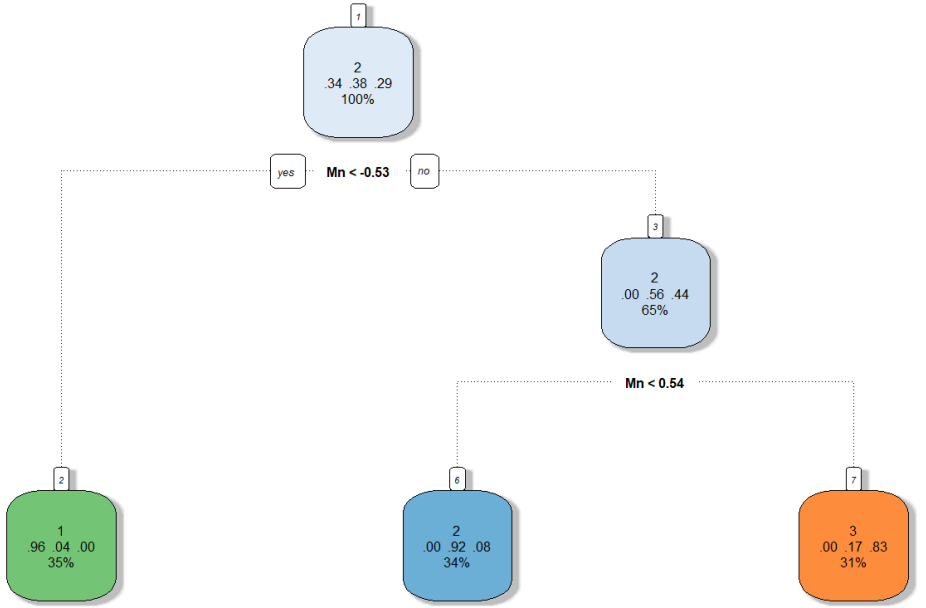
# Breast Cancer Analysis

- 資料處理 : 標準化
- 將資料拆成80%訓練組、20%測試組
- 使用PCA查看是否需做降維動作



# Breast Cancer Analysis

- Decision tree




```
> confus.matrix
     predict
real 1 2 3
   1 5 0 0
   2 1 8 0
   3 0 1 2

> acc
[1] 0.8823529
```

# Breast Cancer Analysis

- Random forest

(R預設的隨機森林為使用多棵 CART樹組成，使用Gini值計算)

trainset

```
Confusion Matrix and Statistics

          Reference
Prediction  1  2  3
         1 18  0  0
         2  2 32  3
         3  0  2 20

Overall Statistics

               Accuracy : 0.9091
                 95% CI : (0.8216, 0.9627)
    No Information Rate : 0.4416
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8583

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            0.9000   0.9412   0.8696
Specificity            1.0000   0.8837   0.9630
Pos Pred Value         1.0000   0.8649   0.9091
Neg Pred Value         0.9661   0.9500   0.9455
Prevalence             0.2597   0.4416   0.2987
Detection Rate         0.2338   0.4156   0.2597
Detection Prevalence   0.2338   0.4805   0.2857
Balanced Accuracy      0.9500   0.9124   0.9163
```

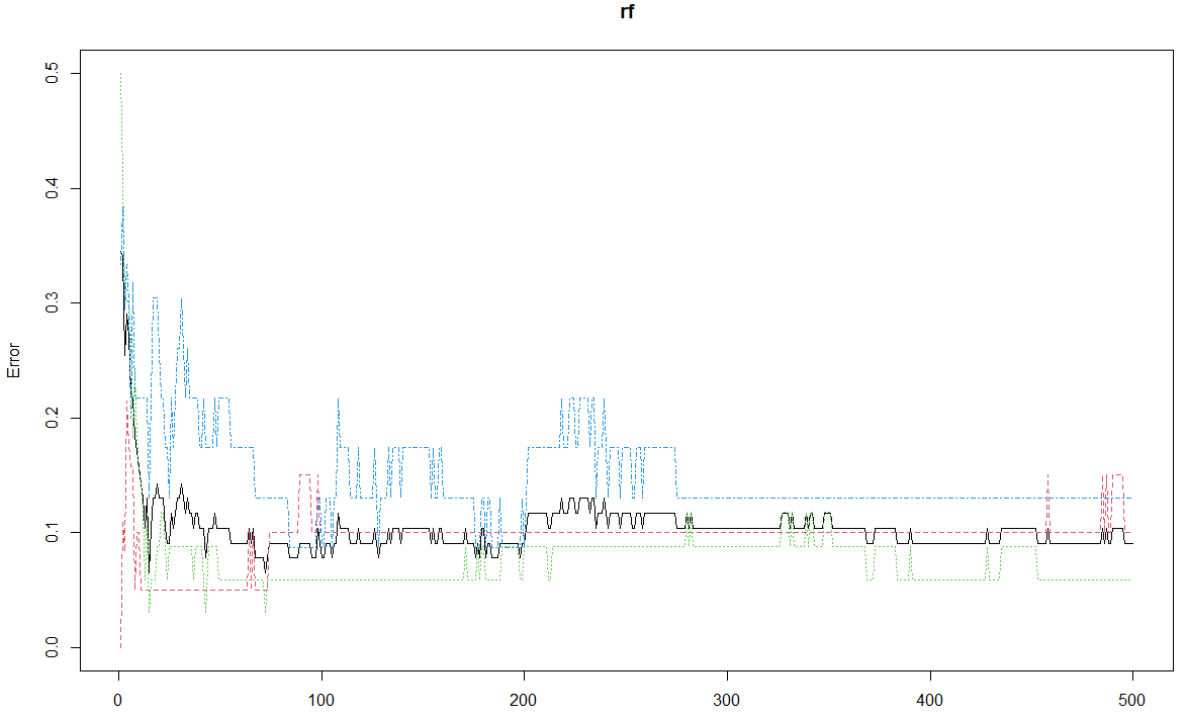
testset

```
Confusion Matrix and Statistics

          Reference
Prediction 1 2 3
         1 5 0 0
         2 0 9 1
         3 0 0 2

Overall Statistics

               Accuracy : 0.9412
                 95% CI : (0.7131, 0.9985)
    No Information Rate : 0.5294
    P-Value [Acc > NIR] : 0.0003248

                  Kappa : 0.8988

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000   1.0000   0.6667
Specificity            1.0000   0.8750   1.0000
Pos Pred Value         1.0000   0.9000   1.0000
Neg Pred Value         1.0000   1.0000   0.9333
Prevalence             0.2941   0.5294   0.1765
Detection Rate         0.2941   0.5294   0.1176
Detection Prevalence   0.2941   0.5882   0.1176
Balanced Accuracy      1.0000   0.9375   0.8333
```

# Breast Cancer Analysis

- Error rate of random forest

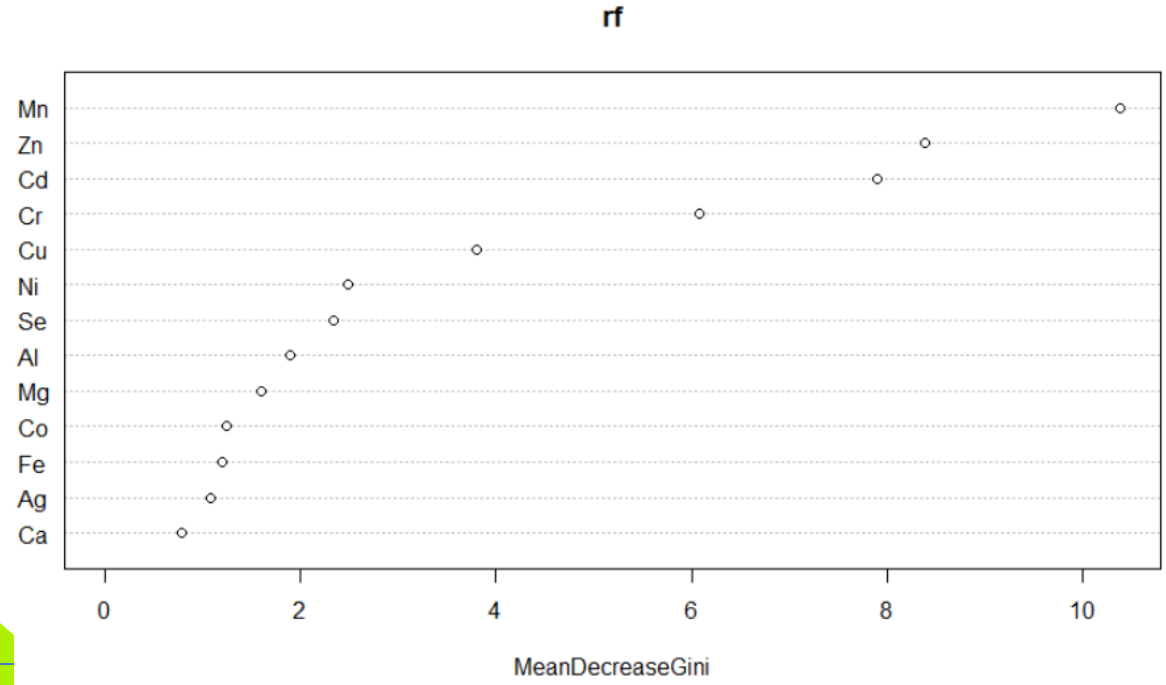
綠：class1
紅：class2
藍：class3
黑：OOB



# Breast Cancer Analysis

◉ 變數重要性圖

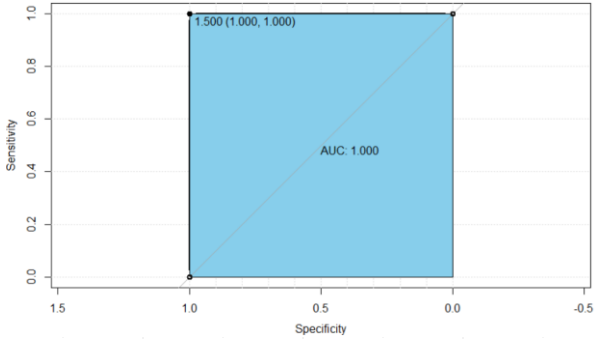
透過 Mean Decrease Gini 來衡量變數重要性指數，表示 Gini 係數減少的平均值



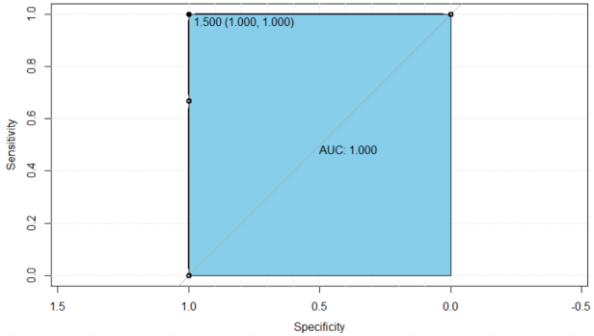
# Breast Cancer Analysis

- ROC curve

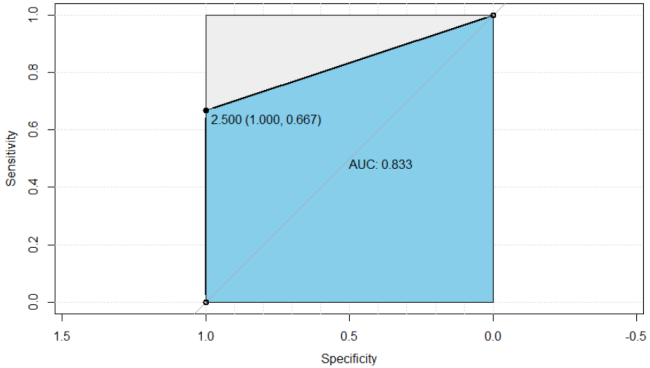
Class1




Class2




Class3



# Breast Cancer Analysis

與其他分類器比較其準確效果

## KNN

```
Confusion Matrix and Statistics

          Reference
Prediction 1 2 3
         1 7 1 0
         2 0 9 1
         3 0 0 2

Overall Statistics

               Accuracy : 0.9
                 95% CI : (0.683, 0.9877)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : 0.0002012

                  Kappa : 0.8319

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3
Sensitivity            1.0000     0.90   0.6667
Specificity            0.9231     0.90   1.0000
Pos Pred Value         0.8750     0.90   1.0000
Neg Pred Value         1.0000     0.90   0.9444
Prevalence             0.3500     0.50   0.1500
Detection Rate         0.3500     0.45   0.1000
Detection Prevalence   0.4000     0.50   0.1000
Balanced Accuracy      0.9615     0.90   0.8333
```

## SVM

```
> cm <- table(x = testset$D1, y = results)
> cm
   y
x   1 2 3
  1 7 0 0
  2 1 8 1
  3 0 0 3
> SVMaccuracy <- sum(diag(cm)) / sum(cm)
> SVMaccuracy
[1] 0.9
```

## Native Bayes

```
> cm <- table(x = testset$D1, y = results)
> cm
   y
x   1 2 3
  1 7 0 0
  2 0 9 1
  3 0 1 2
> naiveBayesaccuracy <- sum(diag(cm)) / sum(cm)
> naiveBayesaccuracy
[1] 0.9
```


Thank you for your attention