1. 介紹決策樹

2. 分類決策樹理論(for y is discrete.)
   y ↓ output

3. 分類決策樹實作

4. 迴歸決策樹理論(for y is continuous)
   y ↓ output

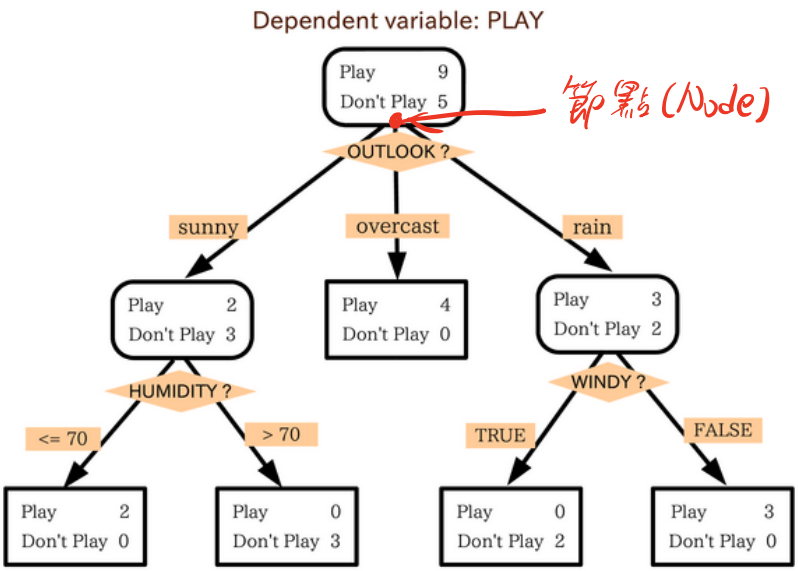
5. 迴歸決策樹實作

6. 隨機森林 ← 下二次.

---

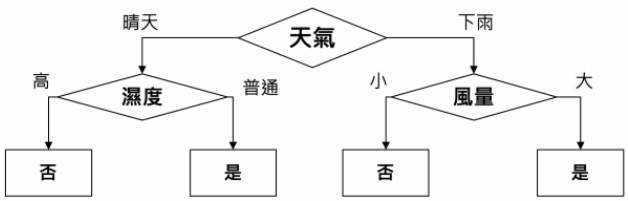
1. 分類決策樹

首先，先看一個例子，要決定是否要玩高爾夫(from wikipedia



另一個例子

天氣	濕度	風量	是否舉行
晴天	高	大	否
陰天	低	小	是



問題來了，分支是如何被決定的？

A：每次分支所獲得的資訊量最大

那如何衡量資訊量的獲得？

A：Gini Index

Gini Index = $1 - \sum p_i^2$

Gini Index 代表的是節點的不純度（Impurity）.

當節點越不純，我們能獲得的資訊越少.

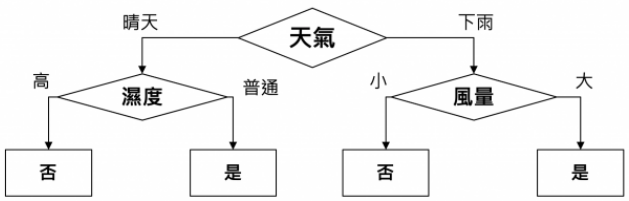
因此，若現在有兩種分類方式，a and b.

$Gini_a > Gini_b$ 則代表 a 方式的 Impurity 較高.

則 a 方式能得到的資訊越少.

因此，我們偏好 b 方法.

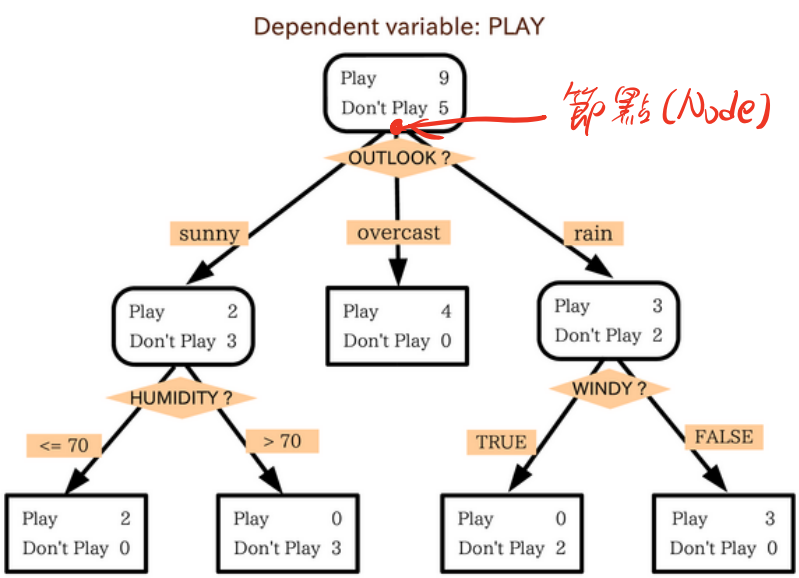
天氣	濕度	風量	是否舉行
晴天	高	大	否
陰天	低	小	是



1. $Gini_{天氣} = 1 - P(晴天)^2 - P(下雨)^2$

$Gini_{濕度} = 1 - P(高)^2 - P(普通)^2$

$Gini_{風量} = 1 - P(大)^2 - P(小)^2$



2. $Gini_{outlook} = 1 - P(sunny)^2 - P(overcast)^2 - P(rain)^2$

$Gini_{humidity} = 1 - P(humidity \leq 70)^2 - P(humidity > 70)^2$

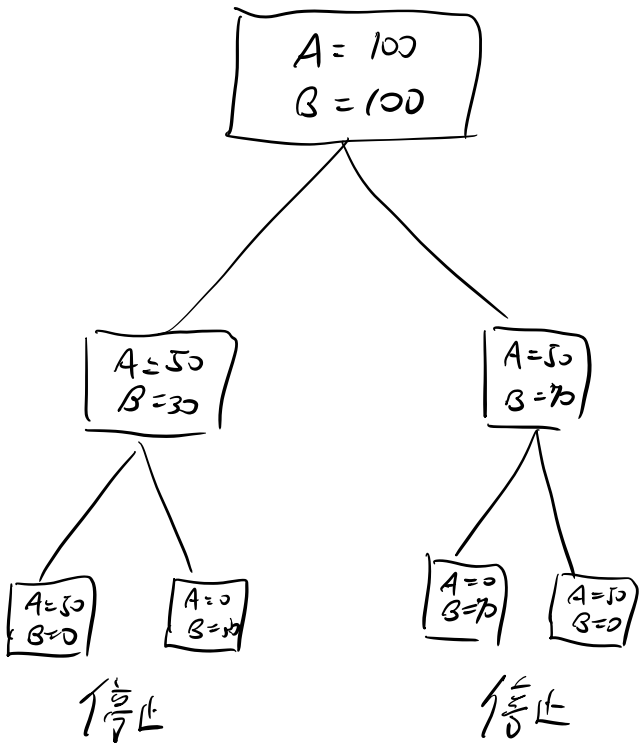
$Gini_{windy} = 1 - P(windy)^2 - P(not\ windy)^2$

何時停止？

當分完數據就結束了！

（當然你也可以加入條件，例如共分類 data < 10）

Ex：

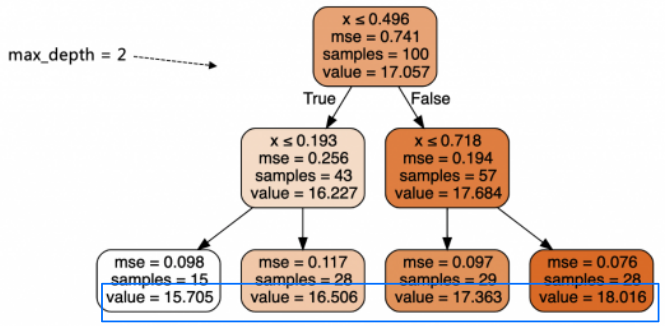


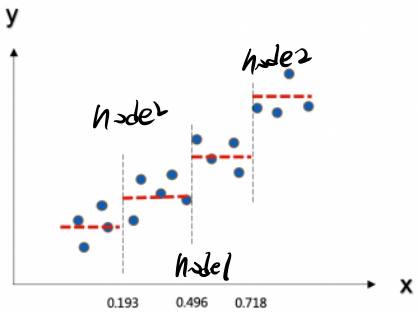
2. 迴歸決策樹

和分類決策樹的不同點在於，y是連續的，而非離散。

舉例而言，y = heigh 時，若要用決策樹，就要用

迴歸決策樹。





預測值

葉節點上，這個預測值為被分到同一類的平均值.

那迴歸決策樹是怎麼分類的呢？

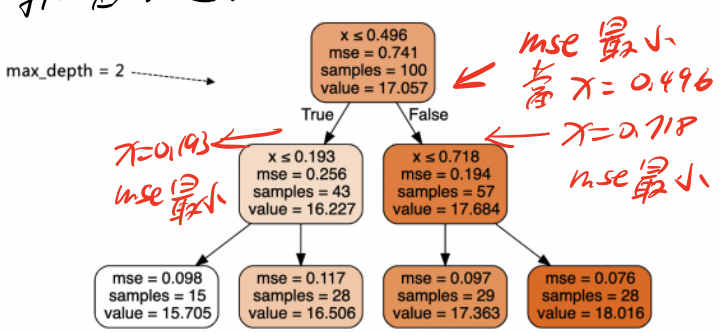
A: 找最小的 mse (和 Regression 想法相同)

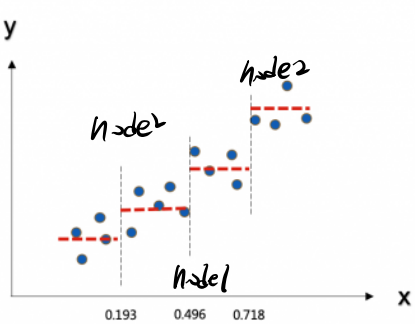
$$MSE = \frac{1}{n}\sum_{i}^{n}(y_i - \hat{y})^2$$

Recall! In Regression, we minimize $LSE = \sum_{i}^{n}(y_i - \hat{y})^2$

因為倍數，因此兩者根本等價.

因此，當決策樹找出某個不的值可以最小化 mse，則就會以這個 x 做分類.





決策樹總結：

Advantage：

易於理解

可以看到模型怎麼判斷的

模型簡單，因此可以快速對大量數據進行處理。

Disadvantage：1. 在連續型資料表現沒有那麼好。

2. 事實上，當某個Y的類別資料特別多，結果會偏向那個類別。

Y=1　　n=50

Y=0　　n=1000

3. 容易 overfitting。

註：X可以不要全放，是我們方便教學才都全放的。