AIoT Lecture 7 Decision Tree and Naive Bayesian

Decision Tree / Entropy, Gini metrics

▼ 0. 前言

1. Google Meet [會議google Meet https://meet.google.com/qjv-fvrx-rka
2. Decision Tree 決策樹
3. Naïve Bayesian 貝式分類器
4. Reference YouTube :machine learning crash course in 10 hours
5. Reference : machine learning A to Z udemy
6. Requires: it 邦， google search ,部落格文章

▼ 1. Naive Bayesian 分類器

1.1 基礎：貝氏定理, 條件機率

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B) * P(B)}{P(A)}$$

Naive Baysian (分類) X : feature $= \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ Y : label

1. 樸素 means X 之間獨立 (非相關) $E[XY] = E[X]E[Y]$ assum

定 $f_{XY}(x,y) = f_X(x) \cdot f_Y(y)$

joint pdf

2. conditional probability

X, Y event $\Rightarrow P(X|Y)$ vs $P(Y|X) = \dfrac{P(X|Y)P(Y)}{P(X)}$

event subset of Sample sample

$P(Y|X) = \dfrac{P(Y \cap X)}{P(X)}$

MECE

Baysian

$$\boxed{\begin{aligned} P(X \cap Y) = P(Y \cap X) &= P(Y|X)P(X) \\ &= P(X|Y)P(Y) \end{aligned}}$$

sample $\begin{cases} 60\text{ 狗 } x \\ 40\text{ 貓 } o \end{cases}$

marginal likelihood

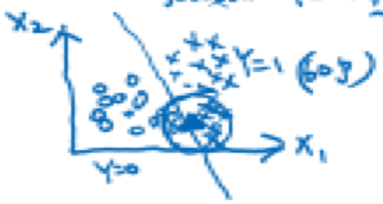
prior probability (先驗) sample

$$\underbrace{P(Y|X)}_{\substack{\text{Postterior} \\ \text{Probability} \\ \text{後驗}}} = \frac{P(X|Y)P(Y)}{P(X)}$$

likelihood

likelihood

feature plan X



$N = 100$ 隻

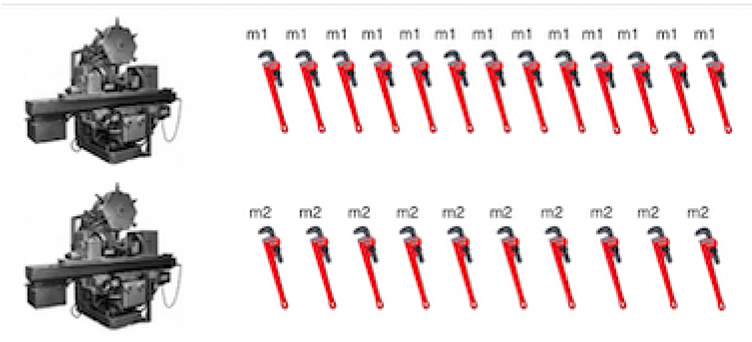
$P(Y|X) =$

$X \begin{cases} p(x,y) : \text{joint prob.} \\ \underline{p(x)} : \text{marginal} = \sum_{\text{all } y} p(x,y) \end{cases}$

$$P(Walks|X) = \frac{P(X|Walks) * P(Walks)}{P(X)}$$

$$Posterior\,Probability = \frac{(Marginal likelihood) \cdot (Prior\,Probability)}{Likelihood}$$

1.2 Naive Bayesian 範例 1



Mach1: 30 wrenches / hr Mach2: 20 wrenches / hr	-> P(Mach1) = 30/50 = 0.6 -> P(Mach2) = 20/50 = 0.4
Out of all produced parts: We can SEE that 1% are defective	-> P(Defect) = 1%
Out of all defective parts: We can SEE that 50% came from mach1 And 50% came from mach2	-> P(Mach1 I Defect) = 50% -> P(Mach2 I Defect) = 50%
Question: What is the probability that a part produced by mach2 is defective = ?	-> P(Defect I Mach2) = ?

Mach1: 30 wrenches / hr Mach2: 20 wrenches / hr Out of all produced parts: We can SEE that 1% are defective Out of all defective parts: We can SEE that 50% came from mach1 And 50% came from mach2 Question: What is the probability that a part produced by mach2 is defective = ?	-> P(Mach2) = 20/50 = 0.4 -> P(Defect) = 1% -> P(Mach2 I Defect) = 50% -> P(Defect I Mach2) = ?

$$P(\text{Defect I Mach2}) = \frac{P(\text{Mach2 I Defect}) * P(\text{Defect})}{P(\text{Mach2})}$$

Question: 壞掉板手的是從機器二做出來的機率？

$$P(Defect|M2) = \frac{0.5 \cdot 0.01}{0.4} = 0.0125 = 1.25\% = \frac{0.5 \cdot 0.01}{0.4} = 0.0125 = 1.25\%$$

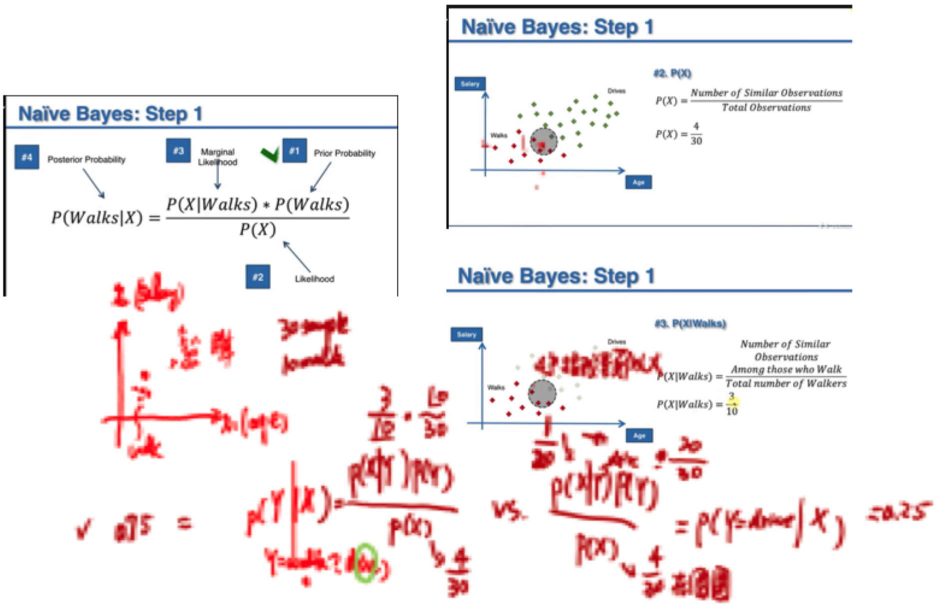
counting 1000個

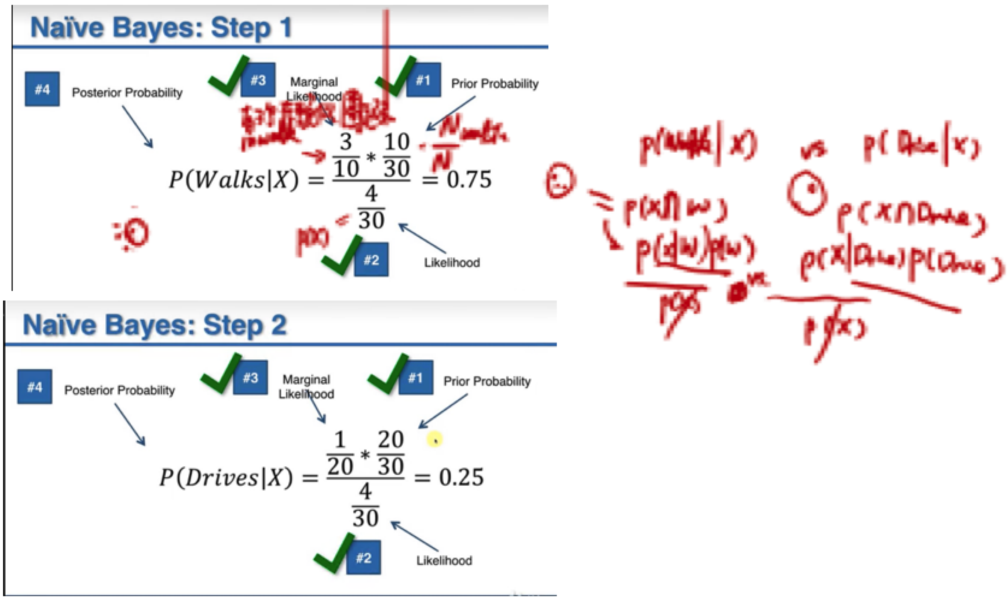
$$P(\text{Defect} \mid \text{Mach2}) = \frac{P(\text{Mach2} \mid \text{Defect}) * P(\text{Defect}) * 1000}{P(\text{Mach2}) * 1000} = 1.25\%$$

Let's look at an example:

- 1000 wrenches
- 400 came from Mach2
- 1% have a defect = 10
- of them 50% came from Mach2 = 5
- % defective parts from Mach2 = 5/400 = 1.25%

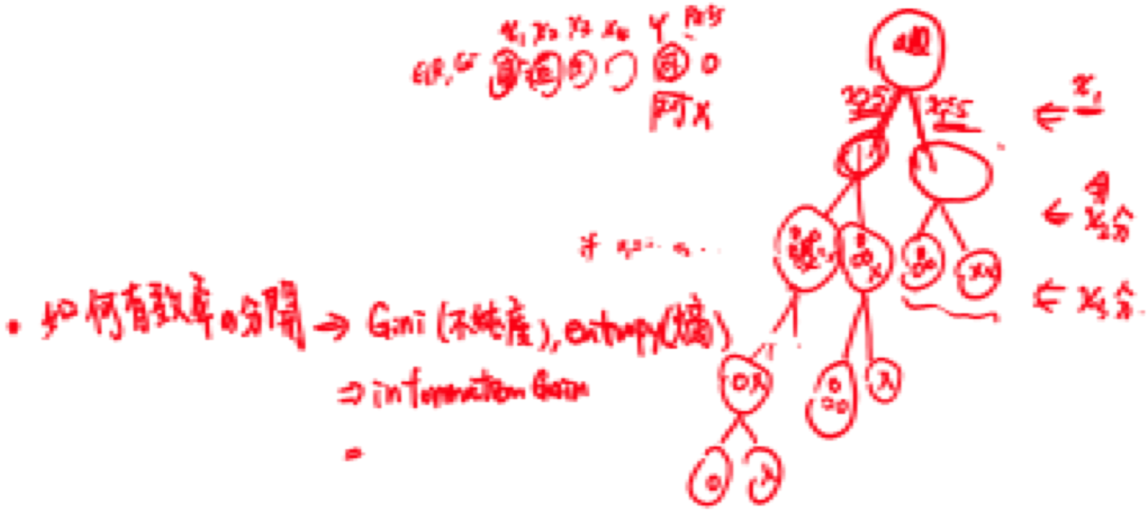
1.3 Naive Bayesian 範例２



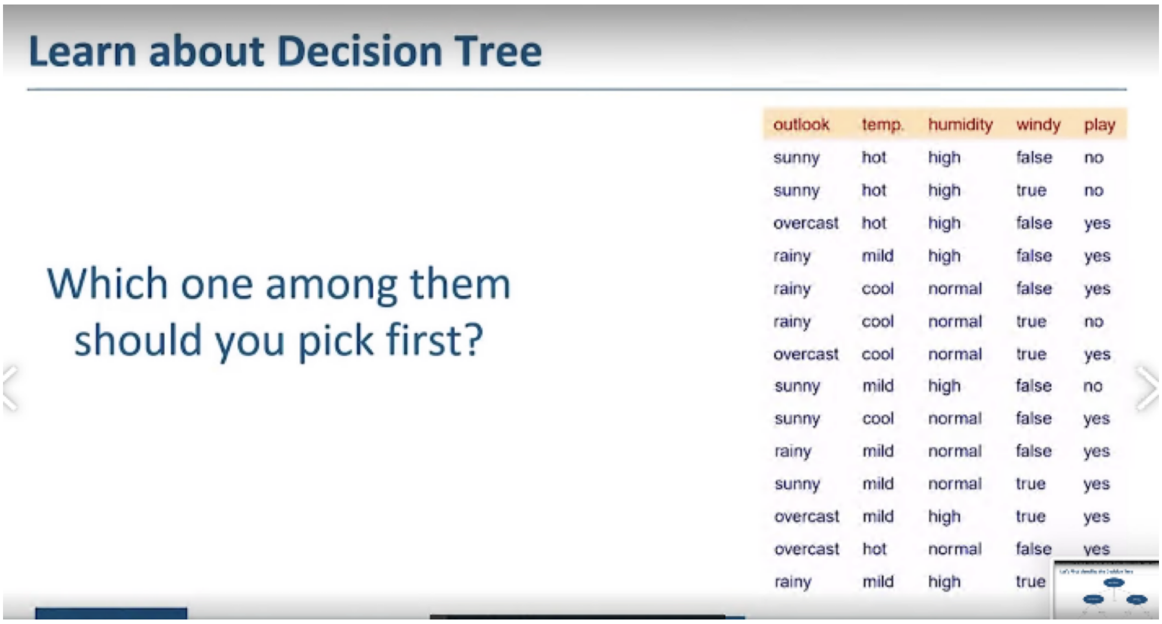


▼ 2. Decision Tree

有名的演算法 CART Classification and Regression Tree



範例



outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	

重點在如何決定 decision criterion 來有效切分=⇒ 使用下列四種: Gini, Entropy, Variance , Chi square 啊啊啊啊

決策樹的混亂評估指標

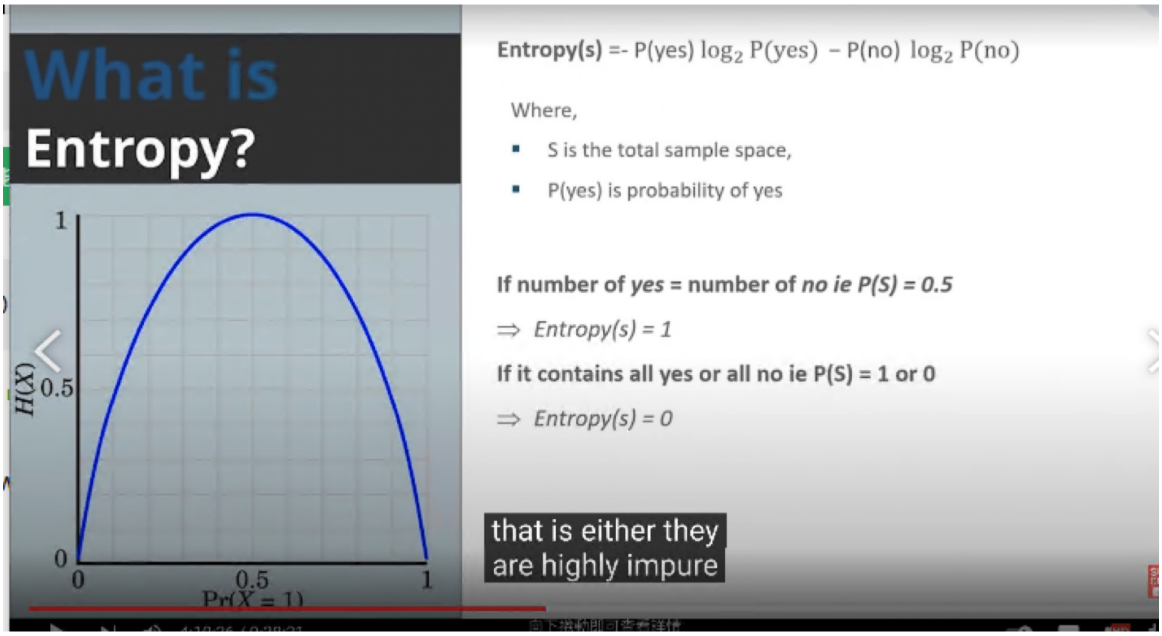
我們需要客觀的標準來決定決策樹的每個分支，因此我們需要有一個評斷的指標來協助我們決策。決策樹演算法可以使用不同的指標來評估分枝的好壞，常見的決策亂度評估指標有 Information gain、Gain ratio、Gini index。我們目標是從訓練資料中找出一套決策規則，讓每一個決策能夠使訊息增益最大化。以上的指標都是在衡量一個序列中的混亂程度，其數值越高代表越混亂。然而在 Sklearn 套件中預設使用 Gini。

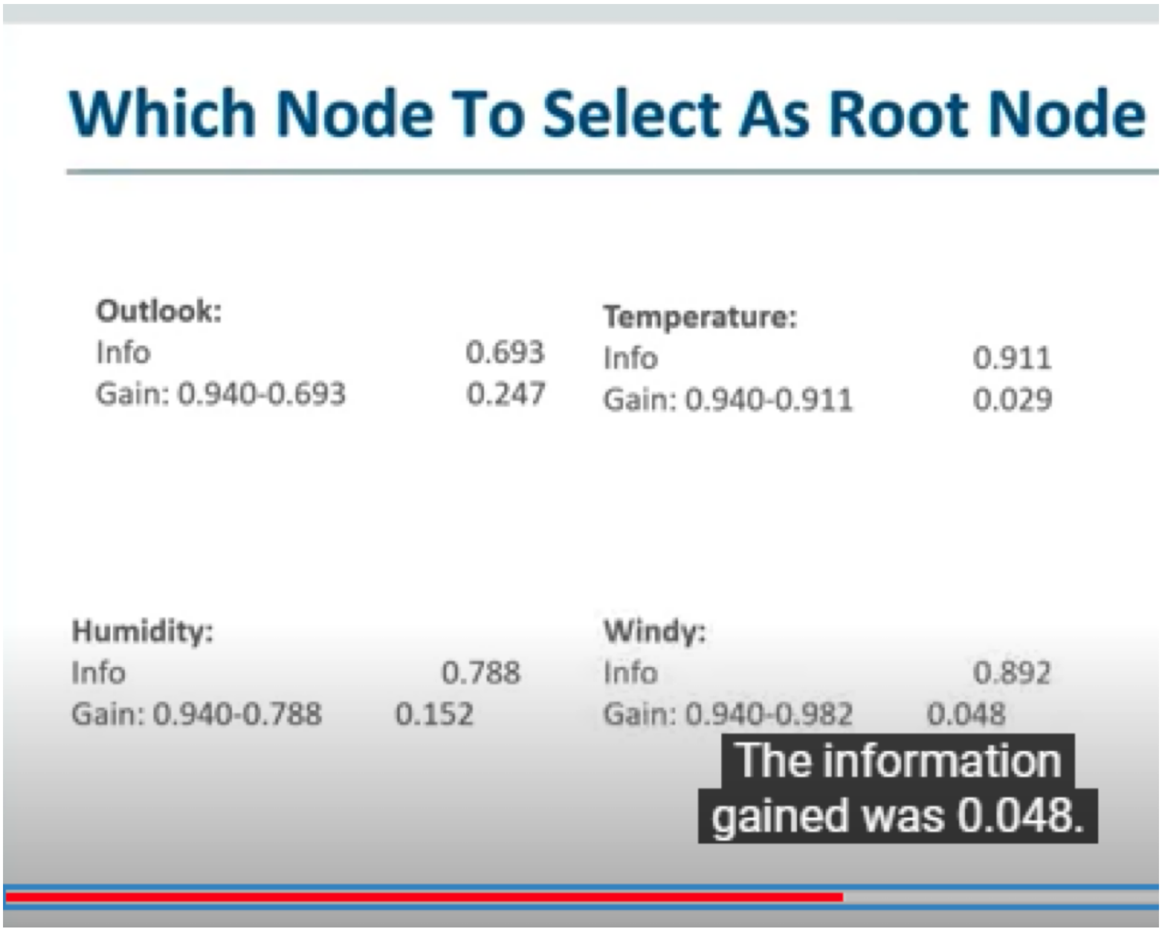
- Information gain (資訊獲利)
- Gain ratio (吉尼獲利)
- Gini index (吉尼係數) = Gini Impurity (吉尼不純度)

- https://ithelp.ithome.com.tw/articles/10271143?sc=pt

用information Gain 來試試, 那什麼是 information gain? 先來了解 entropy, 然後就知道 information gain = 上下分類階層entropy 的差別

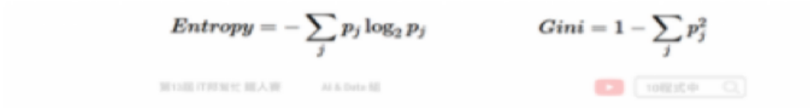
推薦的文章：https://towardsdatascience.com/entropy-and-information-gain-in-decision-trees-c7db67a3a293


What is
Entropy?
1
0.5
0
H(X)
0
0.5
1
Pr(X = 1)
Entropy(s) =- P(yes) $\log_2$ P(yes) − P(no) $\log_2$ P(no)
Where,
S is the total sample space,
P(yes) is probability of yes
If number of yes = number of no ie P(S) = 0.5
⇒ Entropy(s) = 1
If it contains all yes or all no ie P(S) = 1 or 0
⇒ Entropy(s) = 0
that is either they
are highly impure


Which Node To Select As Root Node
Outlook:
Info 0.693
Gain: 0.940-0.693 0.247
Temperature:
Info 0.911
Gain: 0.940-0.911 0.029
Humidity:
Info 0.788
Gain: 0.940-0.788 0.152
Windy:
Info 0.892
Gain: 0.940-0.982 0.048
The information
gained was 0.048.

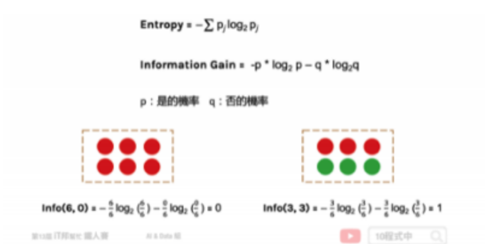
評估分割資訊量

Information Gain 透過從訓練資料找出規則，讓每一個決策能夠使訊息增益最大化。其算法主要是計算熵，因此經由決策樹分割後的資訊量要越小越好。而 Gini 的數值越大代表序列中的資料亂，數值皆為 0~1 之間，其中 0 代表該特徵在序列中是完美的分類。常見的資訊量評估方法有兩種：資訊獲利 (Information Gain) 以及 Gini 不純度 (Gini Impurity)。



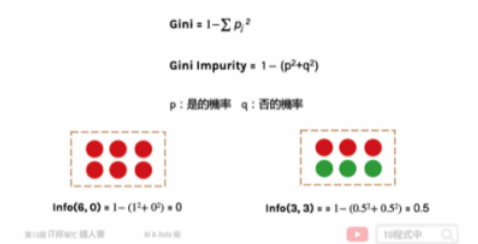
熵 (Entropy)

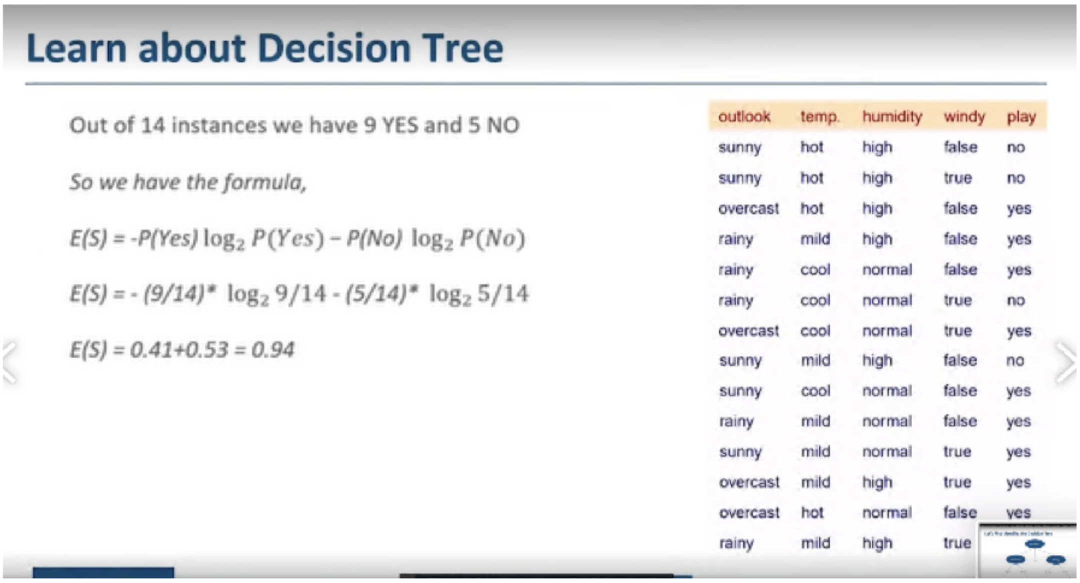
熵 (Entropy) 是計算 Information Gain 的一種方法。在了解 Information Gain 之前要先了解熵是如何被計算出來的。其中在下圖公式中 p 代表是的機率、q 代表否的機率。我們可以從圖中範例很清楚地知道當所有的資料都被分類一致的時候 Entropy 即為 0，當資料各有一半不同時 Entropy 即為 1。



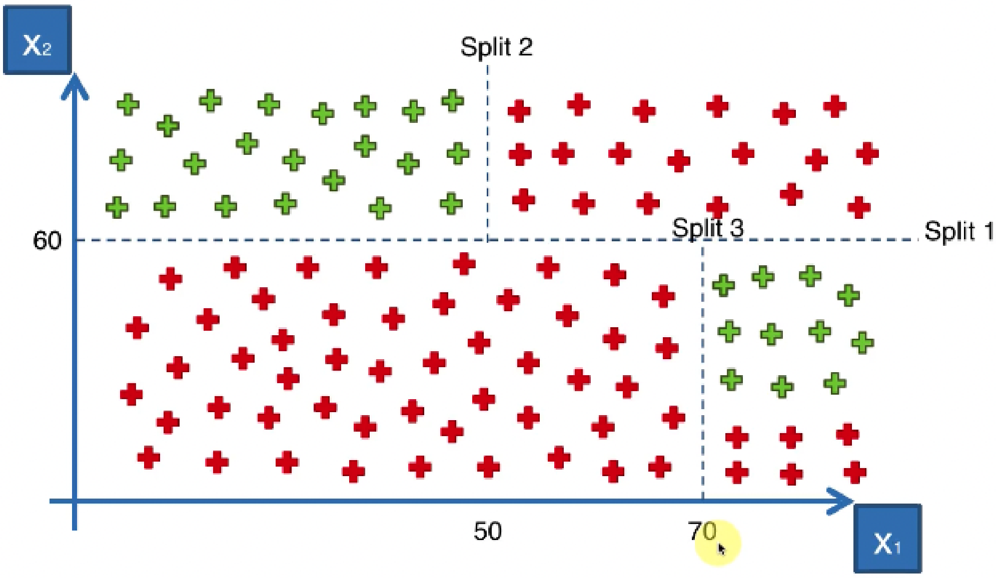
Gini 不純度 (Gini Impurity)

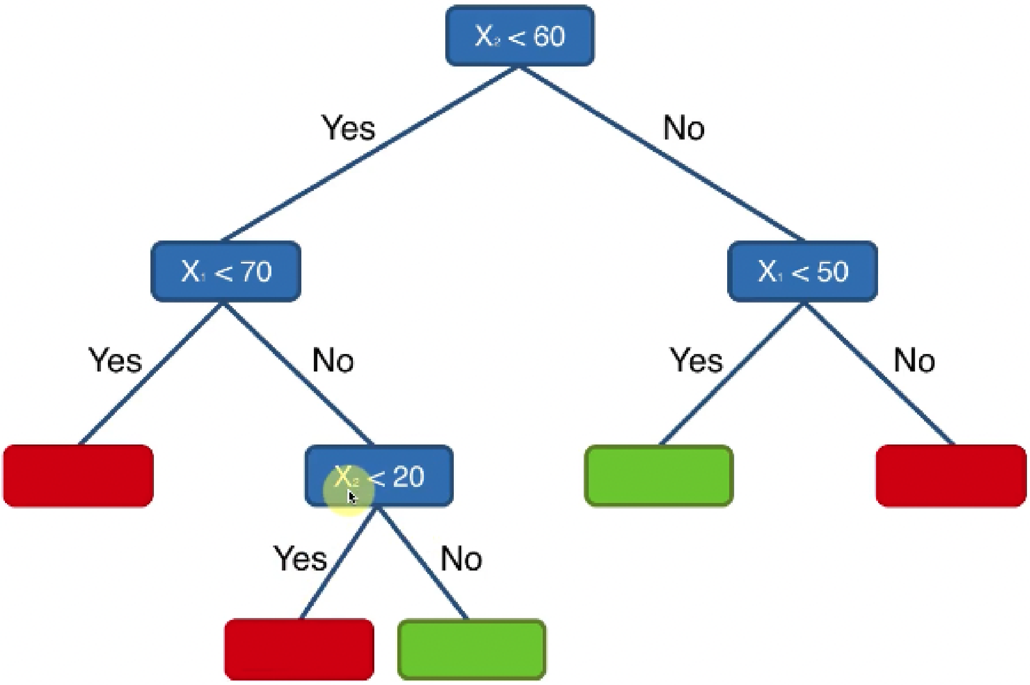
Gini 不純度是另一種亂度的衡量方式，它的數字越大代表序列中的資料越混亂。公式如下所示，其中 p 代表是的機率、q 為代表的機率。我們可以從圖中範例很清楚地知道當所有的資料都被分類一致的時候混亂程度即為 0，當資料各有一半不同時混亂程度即為 0.5。




Learn about Decision Tree
Out of 14 instances we have 9 YES and 5 NO
So we have the formula,
E(S) = -P(Yes) log2 P(Yes) – P(No) log2 P(No)
E(S) = - (9/14)* log2 9/14 - (5/14)* log2 5/14
E(S) = 0.41+0.53 = 0.94
outlook temp. humidity windy play
sunny hot high false no
sunny hot high true no
overcast hot high false yes
rainy mild high false yes
rainy cool normal false yes
rainy cool normal true no
overcast cool normal true yes
sunny mild high false no
sunny cool normal false yes
rainy mild normal false yes
sunny mild normal true yes
overcast mild high true yes
overcast hot normal false yes
rainy mild high true

CART


X2
Split 2
60
Split 3
Split 1
50
70
X1


$X_2 < 60$
Yes
No
$X_1 < 70$
$X_1 < 50$
Yes
No
Yes
No
$X_2 < 20$
Yes
No