決策樹

吳佳諺老師

決策樹

決策樹是一個有方向的無循環圖。
決策樹代表數據的規則。
天氣狀況有晴，雲和雨；氣溫用高低溫度表示；相對濕度用高或一般表示；還有有無風。當然還有是不是真的有騎腳踏車。最終他得到了14列5行的數據表格。

騎腳踏車資料集

天氣	溫度	濕度	有無風	騎腳踏車嗎
太陽	高溫	高	無	不騎
太陽	高溫	高	有	不騎
多雲天	高溫	高	無	騎
下雨	溫和	高	無	騎
下雨	涼	一般	無	騎
下雨	涼	一般	有	不騎
多雲天	涼	一般	有	騎
太陽	溫和	高	無	不騎
太陽	涼	一般	無	騎
下雨	溫和	一般	無	騎
太陽	溫和	一般	有	騎
多雲天	溫和	高	有	騎
多雲天	高溫	一般	無	騎
下雨	溫和	高	有	不騎

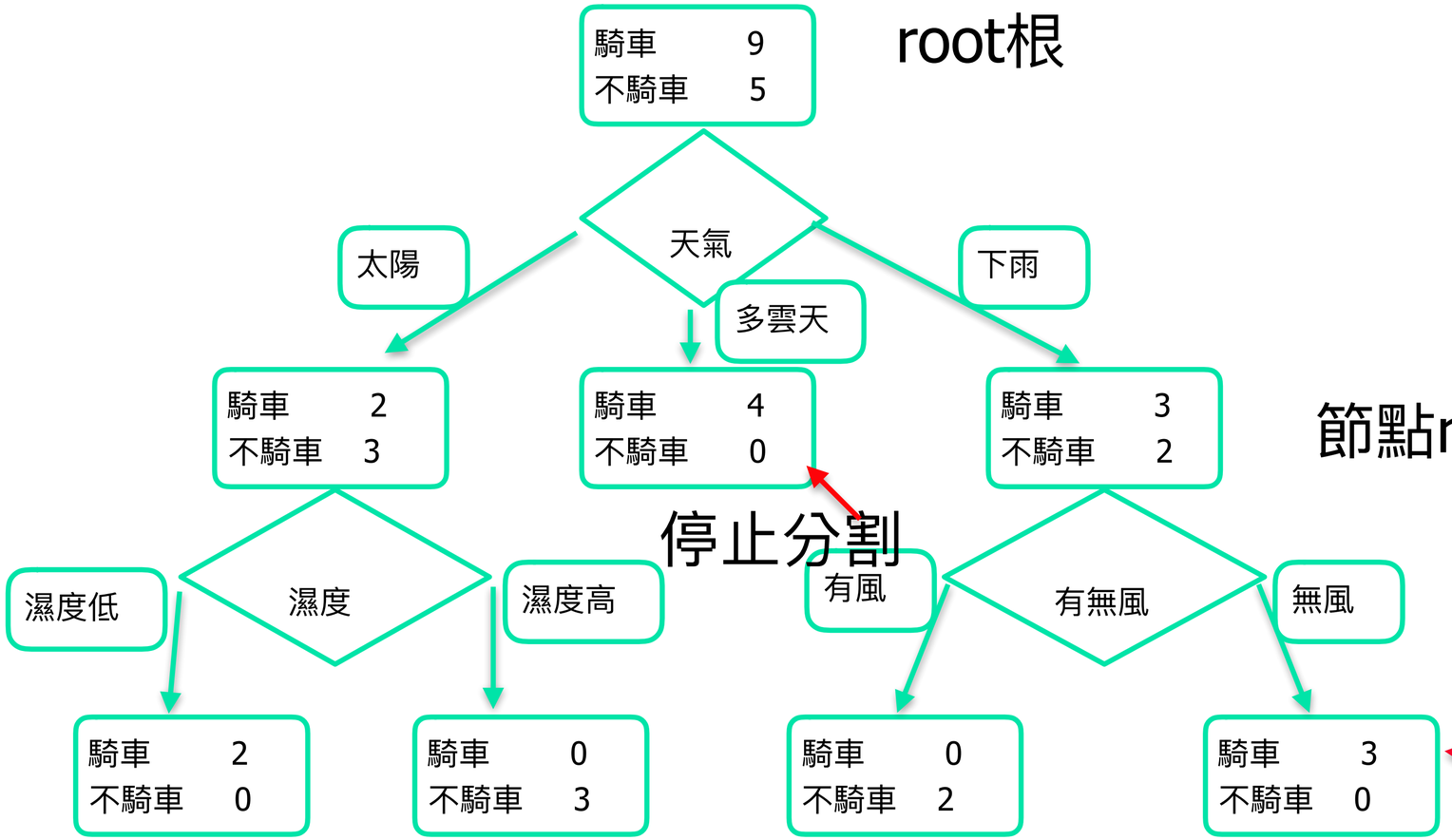
下雨　:低溫　:濕度高　：　無風　：騎腳踏車嗎?

決策樹模型

騎腳踏車天氣的變數範疇被劃分為以下三個組：
太陽，多雲天和下雨天。
如果天氣是多雲，人們總是選擇騎腳踏車，而只有少數在下雨天也會騎。
晴天組的分為兩部分，濕度低與濕度高，不喜歡濕度高的天氣騎腳踏車。
雨天有風的話，就不會有人騎腳踏車。

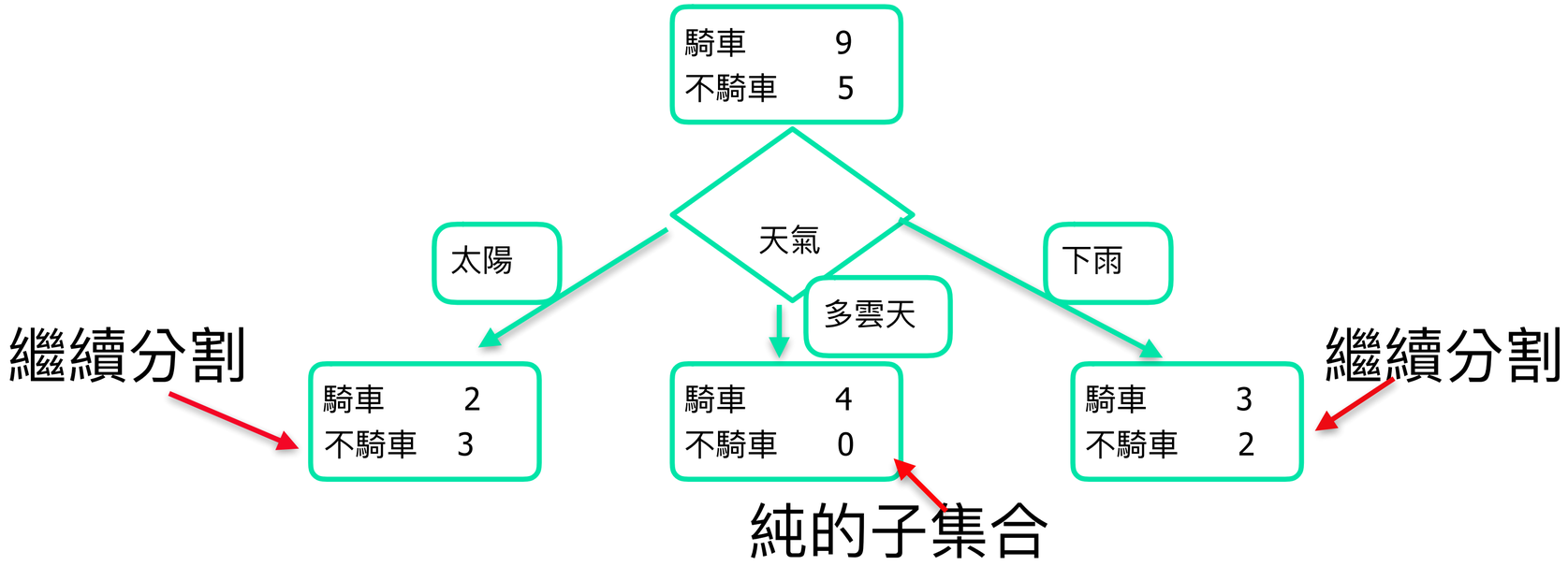
決策樹模型

將資料分割Divide,再結合成一棵樹

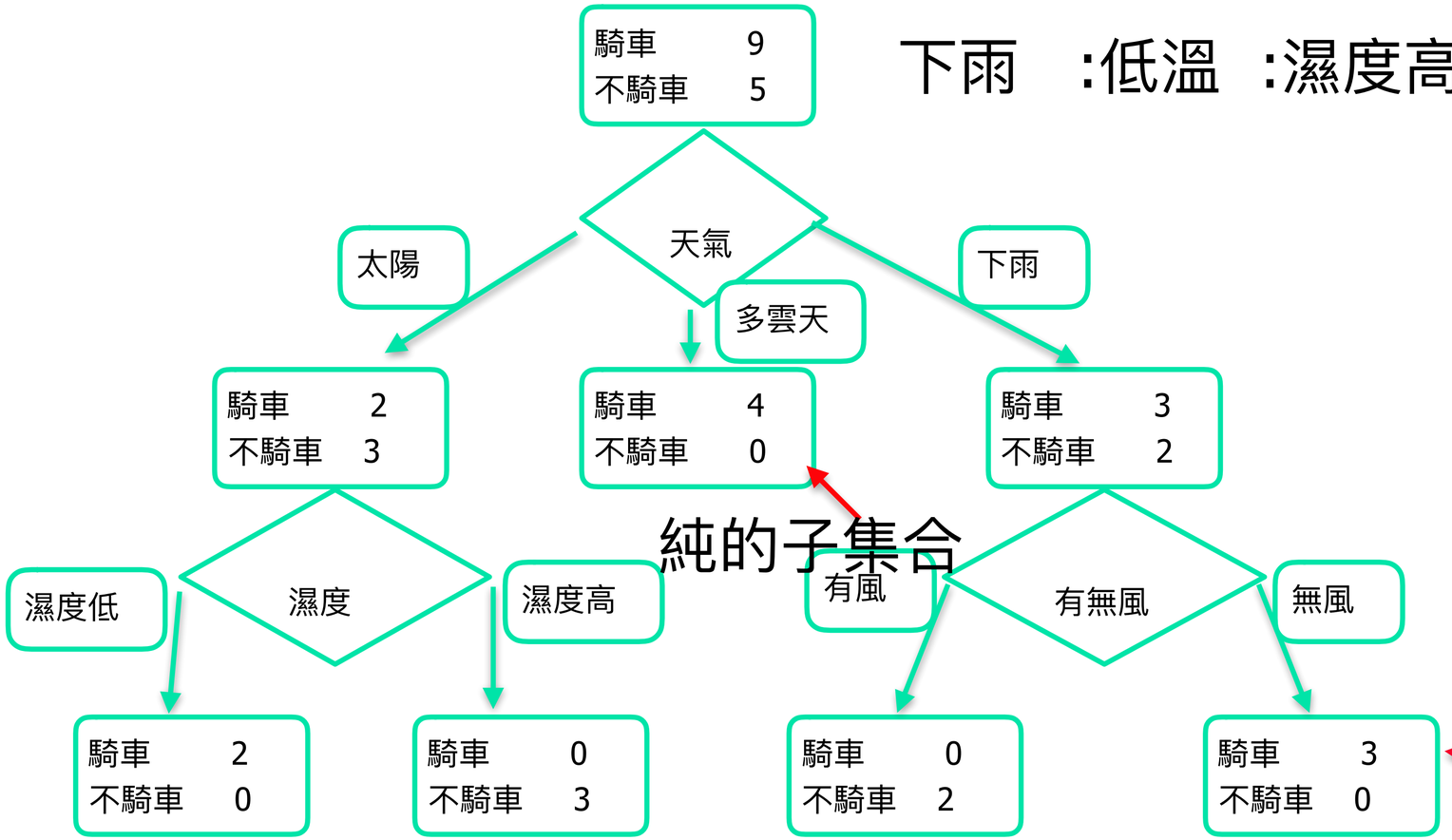


樹葉Leaf

純的子集合可以分成明確的騎車和不騎車



純就是明確,亂度低,也就是熵低

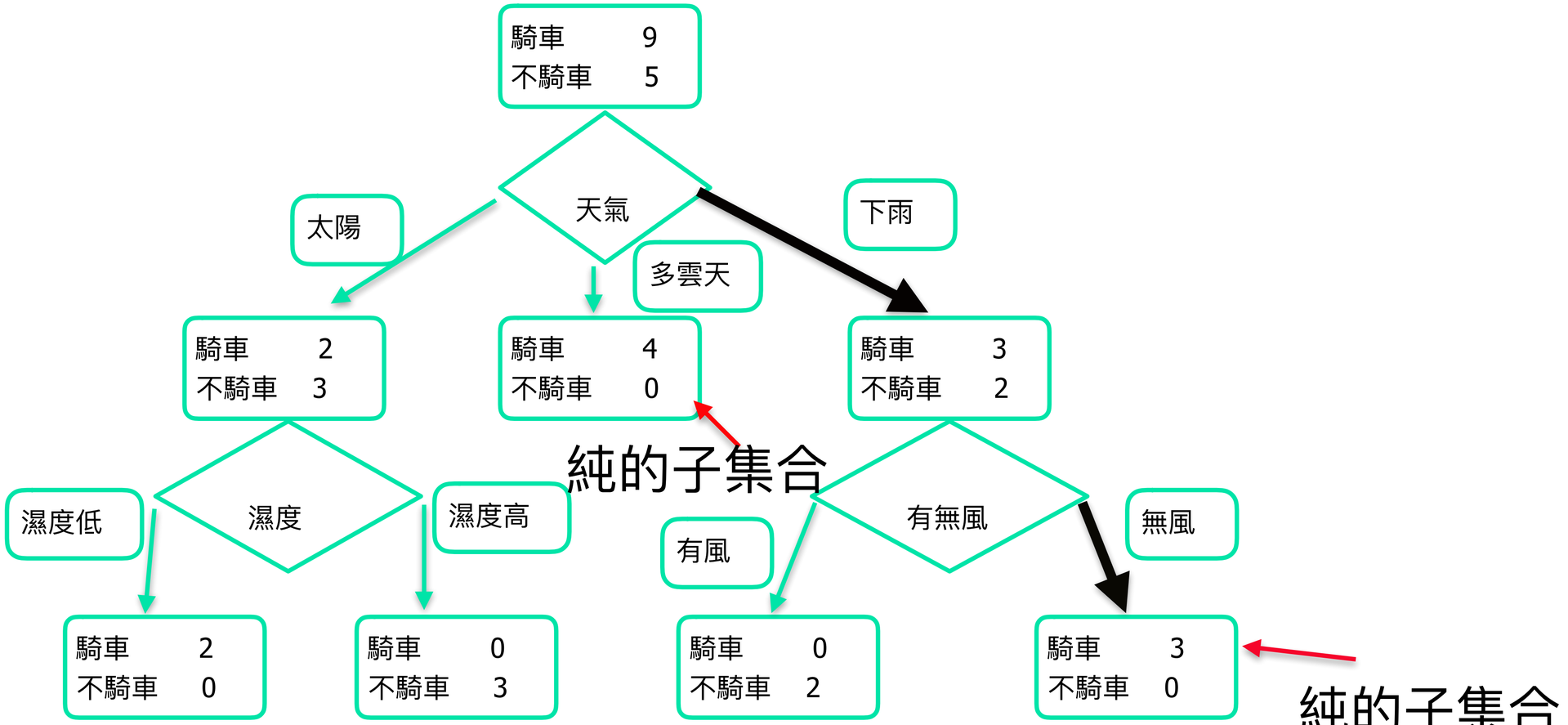


下雨 :低溫 :濕度高 : 無風 : ?

純的子集合

決策樹

下雨　:低溫　:濕度高　：　無風　：　騎車

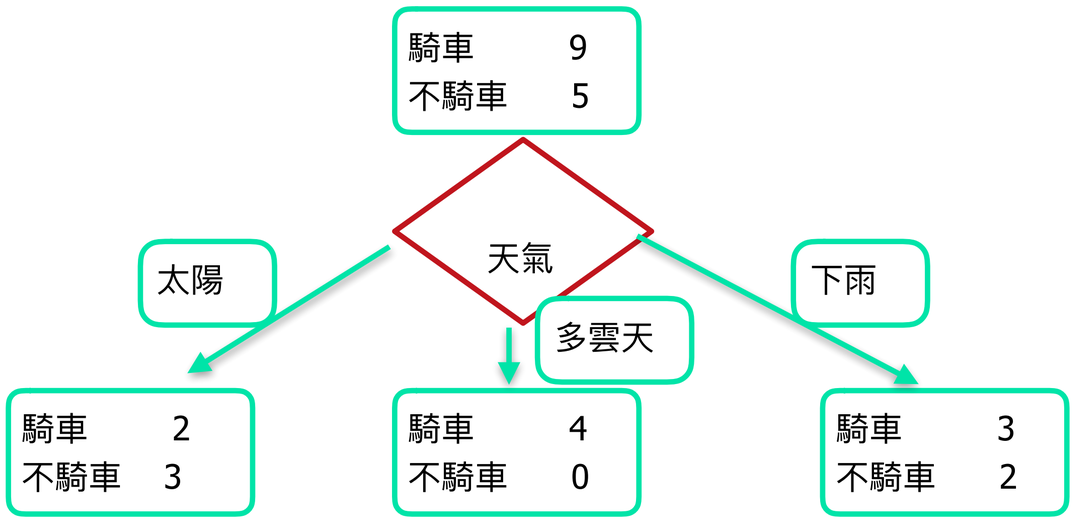


從哪一個屬性開始分割最好

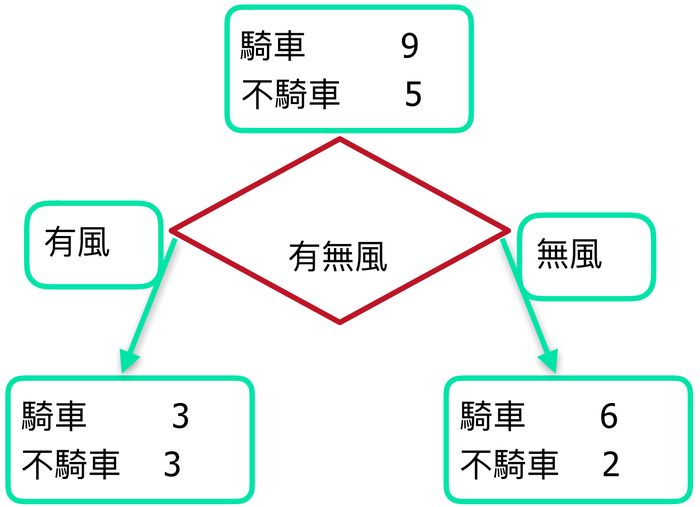
$$Gain(S,\text{天氣})$$
$$=H(S)-\frac{5}{14}H(S_{\text{太陽}})-\frac{5}{14}H(S_{\text{下雨}})$$
$$=0.94-0.6935$$
$$=0.246$$

分割屬性越明確越好

IG=0.246



IG=0.049



H(S)熵Entropy

- S:訓練取樣的資料集
- X:在S中的類別集合
- p(x):在x類別的元素個數對在S資料集的比率

$$H(S) = -\sum_{x \in X} p(x) \log_2 p(x)$$

熵Entropy

- S:訓練取樣的資料集
- p+為正的取樣樣本
- p-為負的取樣樣本
- 假如X屬於S則需要多少位元才能分辨X類別
- 不純度(3 yes / 3 no):需要1位元的資訊才能辨識
- 例如:有無風屬性,有風騎車不純度(3yes/3no)

$$H(S)=-p_+log_2(p_+)-p_-log_2(p_-)$$

$$H(S)=-\frac{3}{6}_{+}log_2(\frac{3}{6})-\frac{3}{6}log_2(\frac{3}{6})=1(位元)$$

熵Entropy

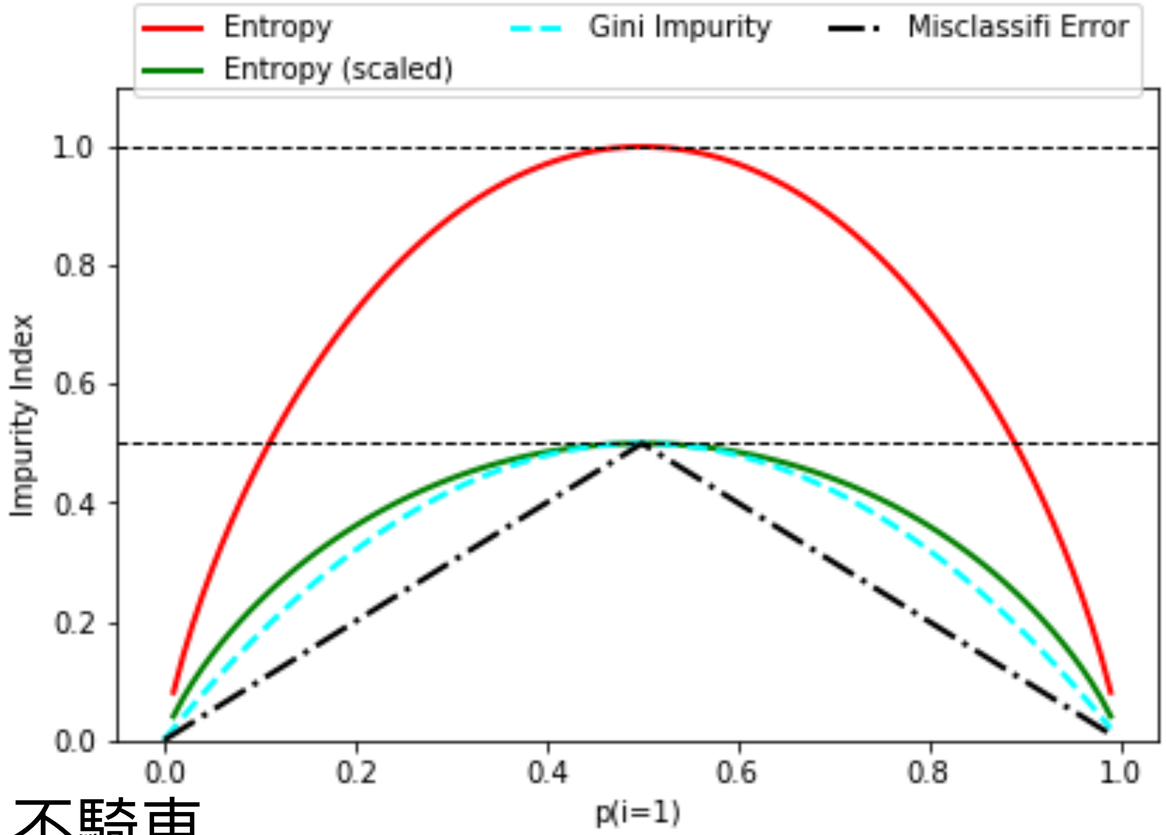
- 純度(6yes / 0 no):辨識度最OK
- 熵Entropy告訴我們取樣資料集合的亂度

$$H(S)=-\frac{6}{6}\log_2\left(\frac{6}{6}\right)-\frac{0}{6}\log_2\left(\frac{0}{6}\right)=0(\text{位元})$$
$$+$$

Entropy熵

- Impurity不純度越低,越容易分辨
- Impurity不純度可用熵Entropy和Gini不純度來代表

多雲天



不騎車

騎車(4,0)

Information Gain資訊增益

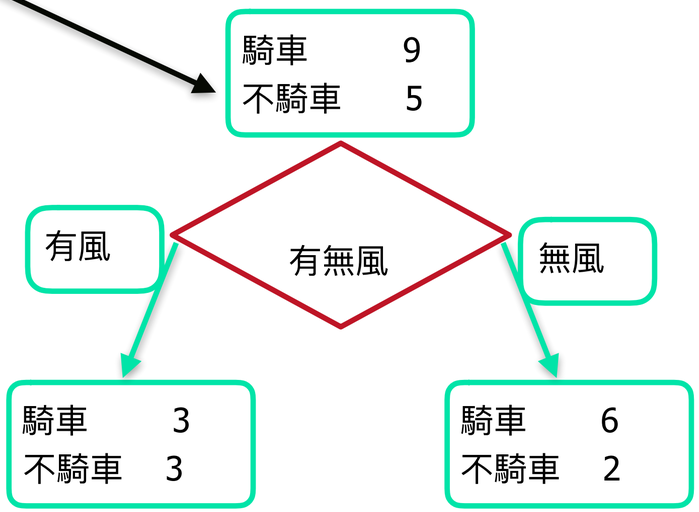
- 希望樣本在越純的集合中容易辨識
- 期望每次節點分割,亂度Entropy能降低
- V是A可能的值,當A為有無風屬性時,其值V為有風或無風
- S是樣本{X(14人)}的集合
- S_v是$X_A=V$的子集合

$S_{無風}$為8,是14人的子集合

$$Gain(S,A)=H(S)-\sum_{V\in Values(A)}\frac{|S_v|}{|S|}H(S_v)$$

Information Gain資訊增益

$$H(S)=\frac{-9}{14}log_2(\frac{9}{14})-\frac{5}{14}log_2(\frac{5}{14})=0.94$$

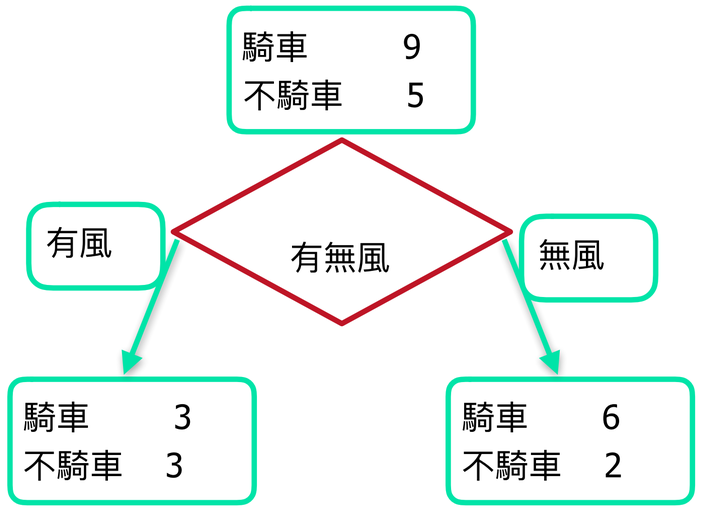


$$H(S)=\frac{-3}{6}log_2(\frac{3}{6})+\frac{-3}{6}log_2(\frac{3}{6})=1.0$$

$$H(S)=\frac{-6}{8}log_2(\frac{6}{8})+\frac{-2}{8}log_2(\frac{2}{8})=0.81$$

Information Gain資訊增益

- 分割有無風的屬性,可以得到Gain資訊增益0.049
- Gain資訊增益越大,表示分辨越明確

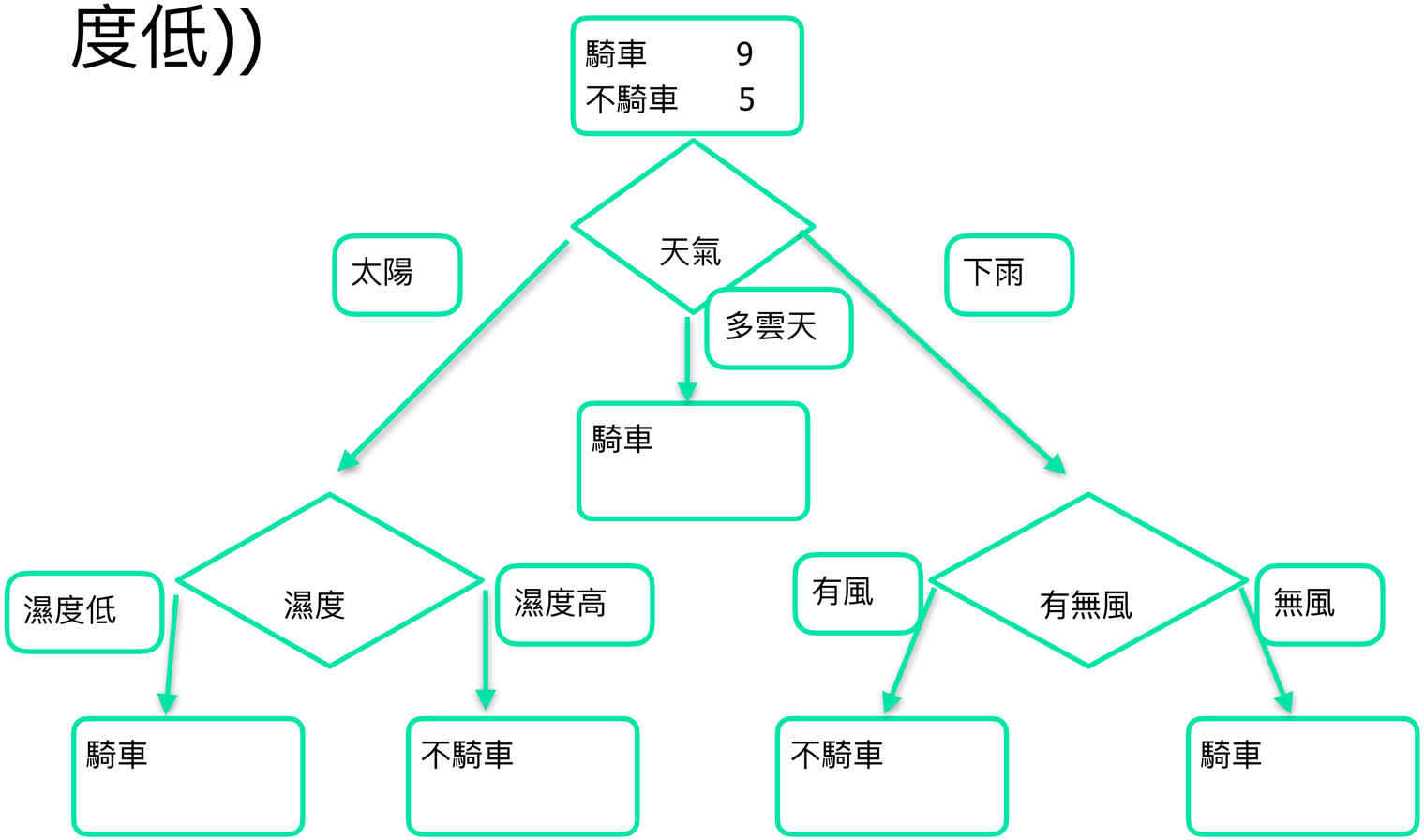


$$Gain(S,\text{有無風})$$
$$=H(S)-\frac{8}{14}H(S_{\text{無風}})-\frac{6}{14}H(S_{\text{有風}})$$
$$=0.94-\frac{8}{14}0.81-\frac{6}{14}1$$
$$=0.049$$

亂度最大

決策樹的規則是易懂的

- 騎車規則:(天氣=多雲天) or ((天氣=下雨天) and (有無風=無風)) or ((天氣=太陽) and (濕度=濕度低))



Information Gain
資訊增益

- H(S):S集合的資訊熵Entropy
- S:目前的資料集
- Ｔ:在S中的子集合,t為有風無風,T為有無風屬性
- p(t):在t類別的元素個數對在S目前資料集的比率
- H(t):t子集的資訊熵
- 資訊增益越大越好 $IG(A,S)$

$$IG(A,S)=H(S)-\sum_{t\in T}p(t)\mathrm{H}(t)$$

- Thanks