decision tree

ID3（Iterative Dichotomiser 3）：

- 使用信息增益作为分割数据的标准。
- 能够处理分类数据。

C4.5：

- 是ID3的后继者，使用信息增益比率来选择特征。
- 能够处理分类和连续数据。
- 对缺失数据有处理能力。

CART（Classification and Regression Trees）：

- 可以用于分类和回归（名字中的R代表回归）。
- 使用基尼不纯度（Gini impurity）或均方误差（Mean Squared Error）作为分割数据的标准。
- 生成的是二叉树。

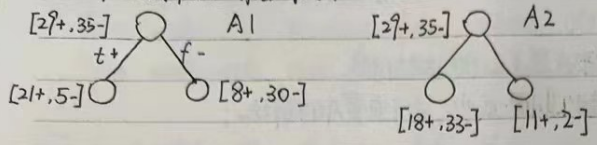
信息熵

$H(A) = -P(A) \cdot log_2 P(A)$

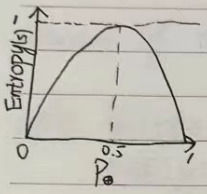
信息增益就是信息熵的差值

No.
Date. / /

Top-Down Induction of Decision Trees:
建立最有效率的决策树.



样本熵. Sample Entropy



S: 训练样本
$p_\oplus$: positive examples 在 S 中占比
$p_\ominus$: negative examples 在 S 中占比

样本S的不纯度熵 $H(S) = -p_\oplus \log_2 p_\oplus - p_\ominus \log_2 p_\ominus$

Entropy $H(X)$ of a random vauable X

$$H(X) = -\sum_{i=1}^{n} P(X=i) \log_2 P(X=i)$$

↓ 期望bits数去encode 随机取出的X值

信息理论: 大多数效率代码需要 $-\log_2 P(X=i)$ bits 去编码信息 $X=i$.
则期望值: $\sum_{i=1}^{n} P(X=i)(-\log_2 P(X=i))$

具体条件下熵 $H(X|Y=v)$ 的 X with given $Y=v$

$$H(X|Y=v) = -\sum_{i=1}^{n} P(X=i|Y=v) \log_2 P(X=i|Y=v)$$

实例

The following data set will be used to learn a decision tree for predicting whether students are lazy (L) or diligent (D) based on their weight (Normal or Underweight), their eye color (Amber, Violet) and the number of eyes they have (2 or 3 or 4).

Weight	Eye Color	Num. Eyes	Output
N	A	2	L
N	V	2	L
N	V	2	L
U	V	3	L
U	V	3	L
U	A	4	D
N	A	4	D
N	V	4	D
U	A	3	D
U	A	3	D

The following numbers may be helpful as you answer this problem without using a calculator: $\log_2 0.1 = -3.32, \log_2 0.2 = -2.32, \log_2 0.3 = -1.73, \log_2 0.4 = -1.32, \log_2 0.5 = -1$.

*You don't need to show the derivation for your answers in this problem.

1. (3 pts) What is the conditional entropy $H(EyeColor|Weight = N)$?

$H(\text{EyeColor}|\text{Weight}=N) = -P(\text{Eye}=A|\text{Weight}=N)\cdot\log_2 P(\text{Eye}=A|\text{Weight}=N)$

$-P(\text{Eye}=V|\text{Weight}=N)\cdot\log_2 P(\text{Eye}=V|\text{Weight}=N) = -0.4\cdot\log_2 0.4 - 0.6\cdot\log_2 0.6 = 0.966$

2. (3 pts) What attribute would the ID3 algorithm choose to use for the root of the tree (before pruning)?

Gain(S, Weight) $= E_1 - \frac{1}{2}E_2 - \frac{1}{2}E_3 =$

Gain(S, Eye Color) $= E_1 - \frac{1}{2}(E_2+E_3)$

Gain(S, Num Eyes) $= E_1 - \frac{3}{10}E_2 - \frac{4}{10}E_3 - \frac{4}{10}E_4 = 0.6$

$E_1 = 1, E_2 = E_3 = E_4 = 0$

3. (4 pts) Draw the full decision tree learned for this data (no pruning).

∵ Gain(S, Num Eyes) is the biggest ∴ Num of Eyes should be the root of the tree.

∵ the output of Num Eyes = 2 is L, the output of Num Eyes = 4 is D

∴ Let's talk about the condition that Num Eyes = 3.

① Use Weight for next Node

S': [2+, 2-]

Weight

N / U

No I_G

$I_G(S', \text{Weight}) = 0$

② Use Eye Color for next Node.

S: [2+, 2-] $E_1 = 1$

Eye Color

A / V

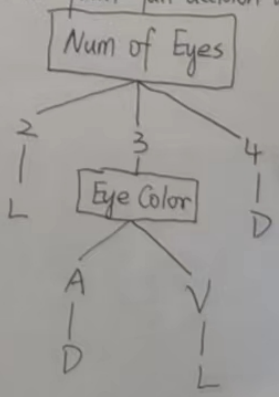
[0+, 2-] [2+, 0-]

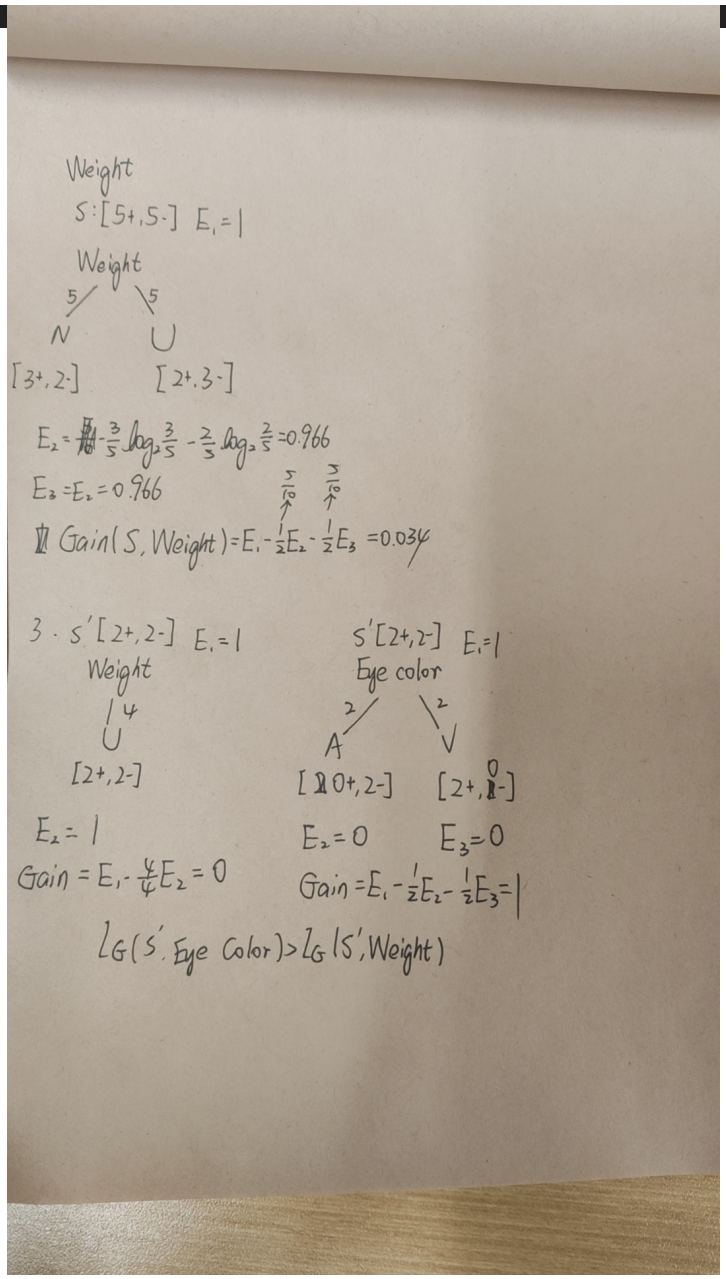
$E_2 = E_3 = 0$

$\therefore I_G(S', \text{Eye Color}) = E_1 - \frac{2}{4}(E_2+E_3) = 1$

$I_G(S', \text{Eye Color}) > I_G(S', \text{Weight})$

∴ the final full decision tree:




Weight
S:[5+,5-] E₁=1
Weight
5
5
N
U
[3+,2-]
[2+,3-]
E₂=-3/5 log₂ 3/5 - 2/5 log₂ 2/5 =0.966
E₃=E₂=0.966
5/10
5/10
Gain(S, Weight)=E₁-1/2E₂-1/2E₃ =0.036
3. S'[2+,2-] E₁=1
Weight
4
[2+,2-]
E₂=1
Gain=E₁-4/4E₂=0
S'[2+,2-] E₁=1
Eye color
2
2
A
V
[2+,0-]
[0+,2-]
E₂=0
E₃=0
Gain=E₁-1/2E₂-1/2E₃=1
IG(S', Eye Color)>IG(S', Weight)