2-1 ID3

2.Mathematical Problems

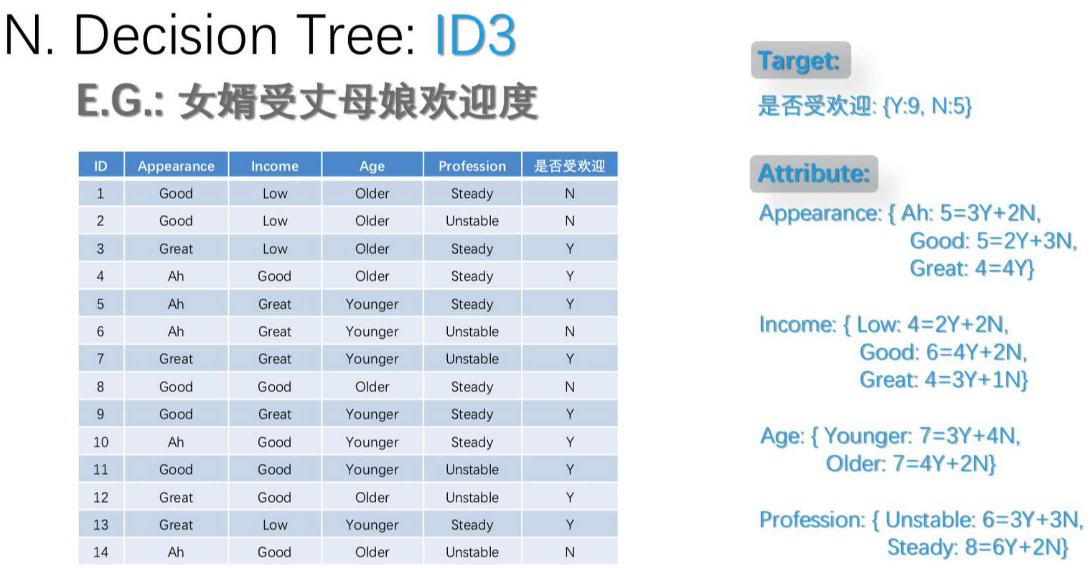
The last topic of our last week is Decision Tree [You can find it in our slides] where we have 3 algorithms (ID3, C4.5 and CART) in which the fundamental one is the ID3 algorithm which was implemented in 1979 initially by Quinlan.

In our slides, we provide a specific example of implementing a decision tree by using ID3 algorithm step by step. The only problem of that is we intentionaly missed some computing details of how to build up the tree.

Therefore your task is to supplyment details of calculation of each node to the tree and provide the evidence why you split a node in such way.

You can follow what we did in slides. And to check whether you set up the correctly or not, you can also find the final tree in our slides.

题目：



N. Decision Tree: ID3

E.G.: 女婿受丈母娘欢迎度

ID	Appearance	Income	Age	Profession	是否受欢迎
1	Good	Low	Older	Steady	N
2	Good	Low	Older	Unstable	N
3	Great	Low	Older	Steady	Y
4	Ah	Good	Older	Steady	Y
5	Ah	Great	Younger	Steady	Y
6	Ah	Great	Younger	Unstable	N
7	Great	Great	Younger	Unstable	Y
8	Good	Good	Older	Steady	N
9	Good	Great	Younger	Steady	Y
10	Ah	Good	Younger	Steady	Y
11	Good	Good	Younger	Unstable	Y
12	Great	Good	Older	Unstable	Y
13	Great	Low	Younger	Steady	Y
14	Ah	Good	Older	Unstable	N

Target:

是否受欢迎: {Y:9, N:5}

Attribute:

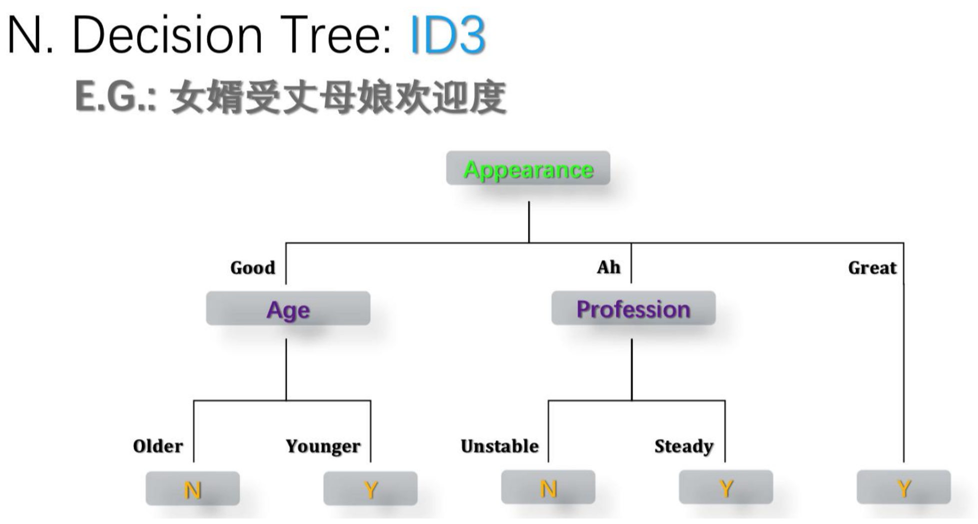
Appearance: { Ah: 5=3Y+2N,
Good: 5=2Y+3N,
Great: 4=4Y}

Income: { Low: 4=2Y+2N,
Good: 6=4Y+2N,
Great: 4=3Y+1N}

Age: { Younger: 7=3Y+4N,
Older: 7=4Y+2N}

Profession: { Unstable: 6=3Y+3N,
Steady: 8=6Y+2N}

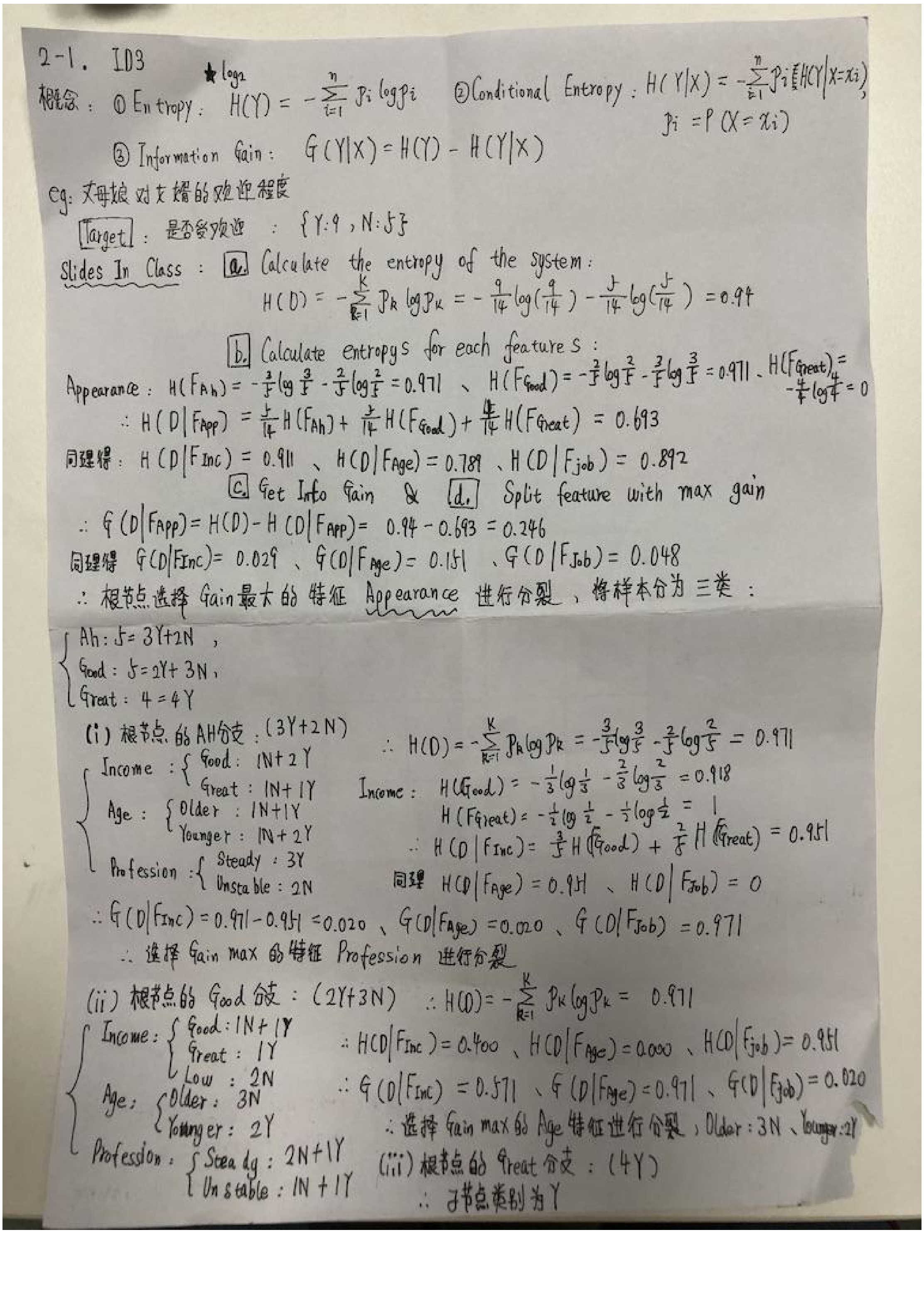
决策树：



N. Decision Tree: ID3

E.G.: 女婿受丈母娘欢迎度

解题过程:



2-1. ID3

★ $\log_2$

概念：① Entropy：$H(Y) = -\sum_{i=1}^{n} p_i \log p_i$ ② Conditional Entropy：$H(Y|X) = -\sum_{i=1}^{n} p_i H(Y|X=x_i)$,

$p_i = P(X=x_i)$

③ Information Gain：$G(Y|X) = H(Y) - H(Y|X)$

eg: 丈母娘对女婿的欢迎程度

$\boxed{\text{Target}}$：是否受欢迎 ：$\{Y:9, N:5\}$

Slides In Class：$\boxed{a.}$ Calculate the entropy of the system：

$$H(D) = -\sum_{k=1}^{K} p_k \log p_k = -\frac{9}{14}\log(\frac{9}{14}) - \frac{5}{14}\log(\frac{5}{14}) = 0.94$$

$\boxed{b.}$ Calculate entropys for each features：

Appearance：$H(F_{Ah}) = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$、$H(F_{Good}) = -\frac{2}{5}\log\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.971$、$H(F_{Great}) = -\frac{4}{4}\log\frac{4}{4} = 0$

$\therefore H(D|F_{App}) = \frac{5}{14}H(F_{Ah}) + \frac{5}{14}H(F_{Good}) + \frac{4}{14}H(F_{Great}) = 0.693$

同理得：$H(D|F_{Inc}) = 0.911$、$H(D|F_{Age}) = 0.789$、$H(D|F_{job}) = 0.892$

$\boxed{c.}$ Get Info Gain & $\boxed{d.}$ Split feature with max gain

$\therefore G(D|F_{App}) = H(D) - H(D|F_{App}) = 0.94 - 0.693 = 0.246$

同理得 $G(D|F_{Inc}) = 0.029$、$G(D|F_{Age}) = 0.151$、$G(D|F_{Job}) = 0.048$

∴ 根节点选择 Gain 最大的特征 Appearance 进行分裂，将样本分为三类：

$$\begin{cases} Ah: 5 = 3Y+2N, \\ Good: 5 = 2Y+3N, \\ Great: 4 = 4Y \end{cases}$$

(i) 根节点的AH分支：(3Y+2N) $\therefore H(D) = -\sum_{k=1}^{K} p_k \log p_k = -\frac{3}{5}\log\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$

Income：{ Good：1N+2Y
Great：1N+1Y

Age：{ Older：1N+1Y
Younger：1N+2Y

Profession：{ Steady：3Y
Unstable：2N

Income：$H(F_{Good}) = -\frac{1}{3}\log\frac{1}{3} - \frac{2}{3}\log\frac{2}{3} = 0.918$

$H(F_{Great}) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$

$\therefore H(D|F_{Inc}) = \frac{3}{5}H(F_{Good}) + \frac{2}{5}H(F_{Great}) = 0.951$

同理 $H(D|F_{Age}) = 0.951$、$H(D|F_{Job}) = 0$

$\therefore G(D|F_{Inc}) = 0.971 - 0.951 = 0.020$、$G(D|F_{Age}) = 0.020$、$G(D|F_{Job}) = 0.971$

∴ 选择 Gain max 的特征 Profession 进行分裂

(ii) 根节点的 Good 分支：(2Y+3N) $\therefore H(D) = -\sum_{k=1}^{K} p_k \log p_k = 0.971$

Income：{ Good：1N+1Y
Great：1Y
Low：2N

Age：{ Older：3N
Younger：2Y

Profession：{ Steady：2N+1Y
Unstable：1N+1Y

$\therefore H(D|F_{Inc}) = 0.400$、$H(D|F_{Age}) = 0.000$、$H(D|F_{Job}) = 0.951$

$\therefore G(D|F_{Inc}) = 0.571$、$G(D|F_{Age}) = 0.971$、$G(D|F_{Job}) = 0.020$

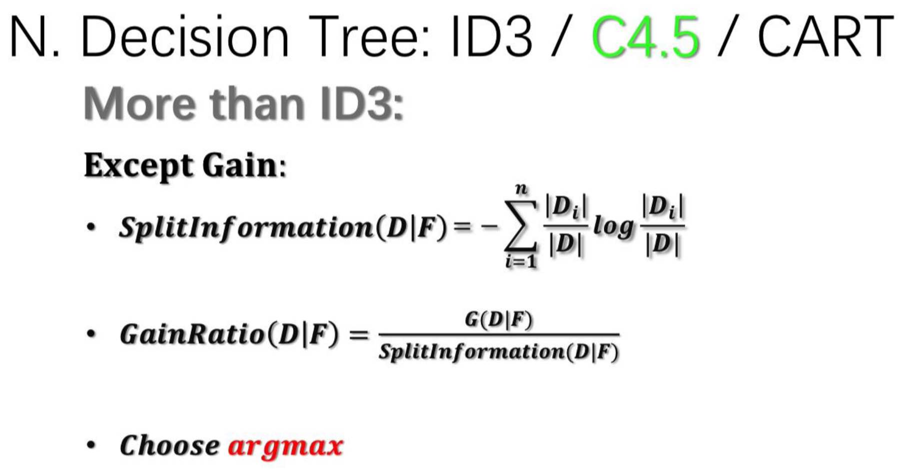
∴ 选择 Gain max 的 Age 特征进行分裂，Older：3N、Younger：2Y

(iii) 根节点的 Great 分支：(4Y)

∴ 子节点类别为Y

2-2 C4.5 & CART

Slides:

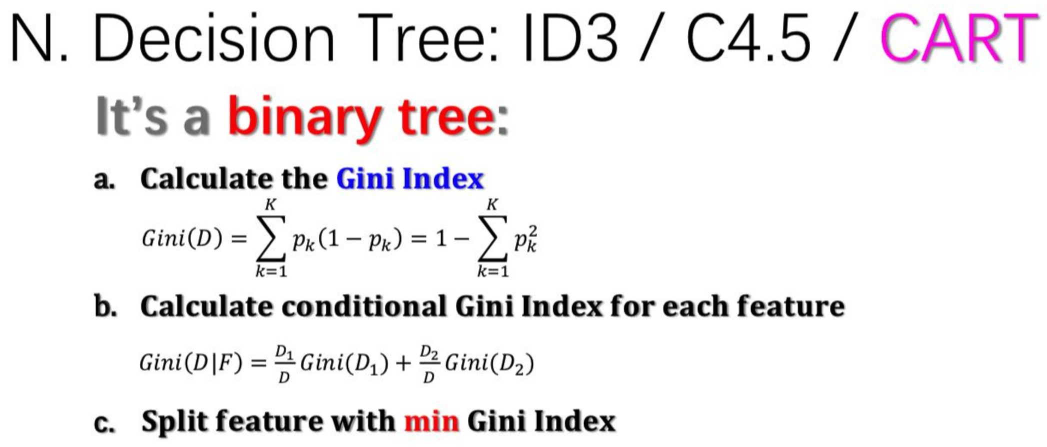


N. Decision Tree: ID3 / C4.5 / CART

More than ID3:

Except Gain:

- $SplitInformation(D|F) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} log \frac{|D_i|}{|D|}$
- $GainRatio(D|F) = \frac{G(D|F)}{SplitInformation(D|F)}$
- *Choose argmax*



N. Decision Tree: ID3 / C4.5 / CART

It's a binary tree:

a. **Calculate the Gini Index**

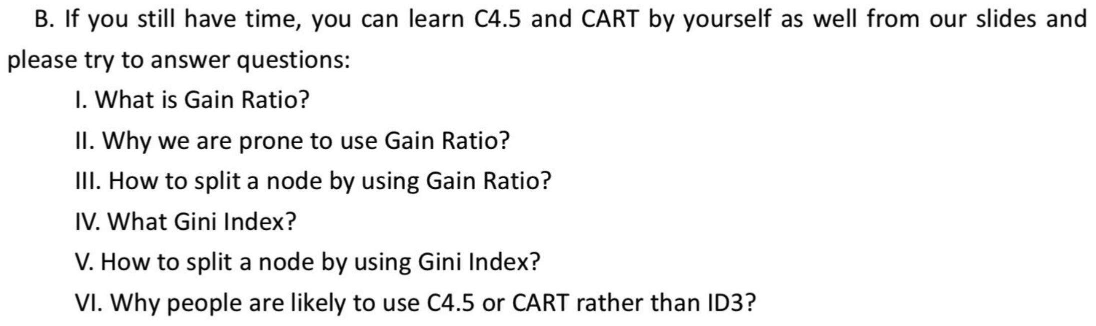
$$Gini(D) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

b. **Calculate conditional Gini Index for each feature**

$$Gini(D|F) = \frac{D_1}{D} Gini(D_1) + \frac{D_2}{D} Gini(D_2)$$

c. **Split feature with min Gini Index**

问题:



B. If you still have time, you can learn C4.5 and CART by yourself as well from our slides and please try to answer questions:

I. What is Gain Ratio?

II. Why we are prone to use Gain Ratio?

III. How to split a node by using Gain Ratio?

IV. What Gini Index?

V. How to split a node by using Gini Index?

VI. Why people are likely to use C4.5 or CART rather than ID3?

答案:

1.

☑ What is Gain Ratio?

特征F对训练集D的信息增益与训练集D关于特征F的值的熵之比。

$$SplitInformation(D|F) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} log \frac{|D_i|}{|D|}$$

$$\text{其中}D_i\text{为样本集}D\text{中}F\text{属性取值为}a_i\text{的子集}$$

$$GainRatio(D|F) = \frac{G(D|F)}{SplitInformation(D|F)}$$

2.

☑ Why we are prone to use Gain Ratio?

信息增益通常对取值数目较多的属性有所偏好。例如使用样本编号作为属性，则分裂后每个子集只有一个样本，则 `Conditional Entropy` 为0，信息增益最大。

信息增益比使用 `SplitInformation` 作为惩罚项，一般取值数目越多，惩罚越大，从而对 `ID3` 的缺点进行校正。

3.

☑ How to split a node by using Gain Ratio?

信息增益比一般对取值数目较多的属性有所偏好。所以一般先从候选属性中选出信息增益高于平均水平的属性，然后再选出信息增益比最高的属性。最后根据该属性的取值，将样本集分裂成若干子集。

4.

☑ What Gini Index?

$$Gini(D) = \sum_{k=1}^{K} p_k(1-p_k) = 1 - \sum_{k=1}^{K} p_k^2$$

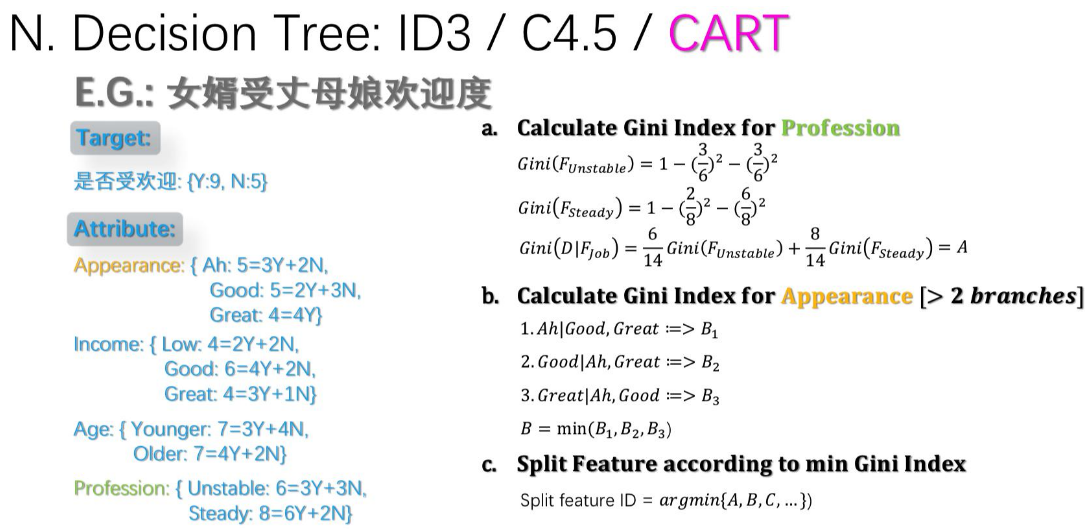
直观来讲， `Gini(D)` 反映了从样本集中随机抽取两个样本，其类别不一致的概率。

5.

☑ How to split a node by using Gini Index?

`CART` 决策树为二叉决策树。设节点的样本集为D，对于每个特征A，对其可能的每个值a，根据A是否等于a将D分割成两个子集D1和D2，计算A=a时的基尼指数。从所有特征A及其所有可能的切分点a中，选择基尼指数最小的特征及切分点。

举例：



N. Decision Tree: ID3 / C4.5 / CART

E.G.: 女婿受丈母娘欢迎度

Target:

是否受欢迎: {Y:9, N:5}

Attribute:

Appearance: { Ah: 5=3Y+2N,
Good: 5=2Y+3N,
Great: 4=4Y}

Income: { Low: 4=2Y+2N,
Good: 6=4Y+2N,
Great: 4=3Y+1N}

Age: { Younger: 7=3Y+4N,
Older: 7=4Y+2N}

Profession: { Unstable: 6=3Y+3N,
Steady: 8=6Y+2N}

a. **Calculate Gini Index for Profession**

$Gini(F_{Unstable}) = 1 - (\frac{3}{6})^2 - (\frac{3}{6})^2$

$Gini(F_{Steady}) = 1 - (\frac{2}{8})^2 - (\frac{6}{8})^2$

$Gini(D|F_{Job}) = \frac{6}{14} Gini(F_{Unstable}) + \frac{8}{14} Gini(F_{Steady}) = A$

b. **Calculate Gini Index for Appearance [> 2 *branches*]**

$1. Ah|Good, Great \coloneqq> B_1$

$2. Good|Ah, Great \coloneqq> B_2$

$3. Great|Ah, Good \coloneqq> B_3$

$B = \min(B_1, B_2, B_3)$

c. **Split Feature according to min Gini Index**

Split feature ID = $argmin\{A, B, C, \ldots\})$

6.

☑ Why people are likely to use C4.5 or CART rather than ID3?

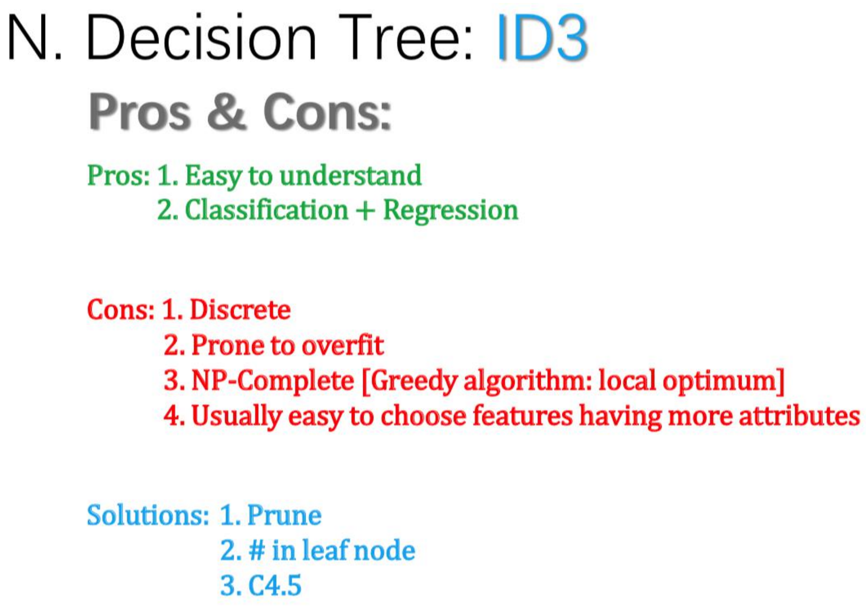
`C4.5` 和 `CART` 克服了 `ID3` 的一些缺点。

`ID3` 对取值数目较多的属性有所偏好， `ID3` 引入 `SplitInformation` 进行校正。

`ID3` 无法处理属性值为连续值以及属性值缺失的情况。 `C4.5` 采用二分法对连续属性进行离散化，从而支持连续属性。通过引入样本权重，对信息增益的表达式进行推广，从而支持缺失值问题。

`CART` 同样可以处理属性连续值和属性值缺失的情况。同时还可以处理回归问题、异常点检测。

ID3 的缺点：



N. Decision Tree: ID3

Pros & Cons:

Pros: 1. Easy to understand
2. Classification + Regression

Cons: 1. Discrete
2. Prone to overfit
3. NP-Complete [Greedy algorithm: local optimum]
4. Usually easy to choose features having more attributes

Solutions: 1. Prune
2. # in leaf node
3. C4.5