Decision Tree

ID3：根据信息增益生成决策树

信息熵就是所有可能发生的事件的信息量的期望$H(Y) = -\sum_{i=1}^{n} P(y_i)\log P(y_i)$

条件熵，在 X 给定条件下，Y 的条件概率分布的熵对 X 的数学期望：

$$H(Y|X) = \sum_{x\in X} P(x)H(Y|X=x)$$
$$= -\sum_{x\in X} P(x)\sum_{y\in Y} P(y|x)\log P(y|x)$$
$$= -\sum_{x\in X}\sum_{y\in Y} P(x,y)\log P(y|x)$$

信息增益$Gain(Y,X) = H(Y) - H(Y|X)$：

根据如下数据集生成决策树

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

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=6Y+1N, Older: 7=3Y+4N}

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

1. Step1: 计算总的 Entropy

$$H(D = 受欢迎) = -\frac{9}{14}log\frac{9}{14} - \frac{5}{14}log\frac{5}{14} = 0.94$$

2. Step2: 计算每个特征的 Entropy

Appearance: $H(App = Great) = -\frac{4}{4}\log\frac{4}{4} = 0$

$$H(App = Good) = -\frac{2}{5}log\frac{2}{5} - \frac{3}{5}log\frac{3}{5} = 0.971$$

$$H(App = Ah) = -\frac{3}{5}log\frac{3}{5} - \frac{2}{5}log\frac{2}{5} = 0.971$$

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

Income: $H(Inc = Great) = -\frac{1}{4}log\frac{1}{4} - \frac{3}{4}log\frac{3}{4} = 0.811$

$$H(Inc = Good) = -\frac{4}{6}log\frac{4}{6} - \frac{2}{6}log\frac{2}{6} = 0.918$$

$$H(Inc = Low) = -\frac{2}{4}log\frac{2}{4} - \frac{2}{4}log\frac{2}{4} = 1$$

$$H(D|F_{Income}) = \frac{4}{14} * H(Inc = Great) + \frac{6}{14} * H(Inc = Good) + \frac{4}{14} * H(Inc = Low) = 0.911$$

Age: $H(Age = Younger) = -\frac{6}{7}log\frac{6}{7} - \frac{1}{7}log\frac{1}{7} = 0.592$

$$H(Age = Older) = -\frac{4}{7}log\frac{4}{7} - \frac{3}{7}log\frac{3}{7} = 0.985$$

$$H(D|F_{Age}) = \frac{7}{14} * H(Age = Younger) + \frac{7}{14} * H(Age = Older) = 0.789$$

Profession: $H(Prof = Steady) = -\frac{6}{8}log\frac{6}{8} - \frac{2}{8}log\frac{2}{8} = 0.811$

$$H(Prof = Unstable) = -\frac{3}{6}log\frac{3}{6} - \frac{3}{6}log\frac{3}{6} = 1$$

$$H(D|F_{Prof}) = \frac{8}{14} * H(Prof = Steady) + \frac{6}{14} * H(Prof = Unstable) = 0.892$$

3. Step3: 计算每个特征的信息增益，取增益最大的作为根节点。

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

$G(D|F_{Inc}) = H(D) - H(D|F_{Inc}) = 0.94 - 0.911 = 0.029$

$G(D|F_{Age}) = H(D) - H(D|F_{Age}) = 0.94 - 0.789 = 0.151$

$G(D|F_{Prof}) = H(D) - H(D|F_{Prof}) = 0.94 - 0.892 = 0.048$

可见，Appearance 的信息增益最大，取 Appearance 作为根结点。将原数据集分为如下 3 个子集：

D1(App=Great)

ID	Appearance	Income	Age	Profession	是否受欢迎
3	Great	Low	Older	Steady	Y
7	Great	Great	Younger	Unstable	Y
12	Great	Good	Older	Unstable	Y
13	Great	Low	Younger	Steady	Y

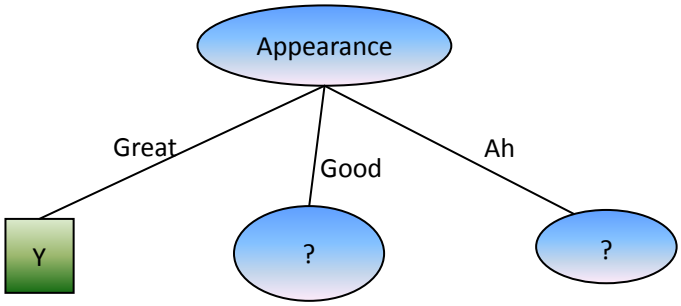
D2(App=Good)

ID	Appearance	Income	Age	Profession	是否受欢迎
1	Good	Low	Older	Steady	N
2	Good	Low	Older	Unstable	N
8	Good	Good	Older	Steady	N
9	Good	Great	Younger	Steady	Y
11	Good	Good	Younger	Unstable	Y

D3(App=Ah)

ID	Appearance	Income	Age	Profession	是否受欢迎
4	Ah	Good	Older	Steady	Y
5	Ah	Great	Younger	Steady	Y
6	Ah	Great	Younger	Unstable	N
10	Ah	Good	Younger	Steady	Y
14	Ah	Good	Older	Unstable	N

当 Appearance=Great 时，即 D1 中只有同一类的样本，所以它为一个叶结点，结点的类标记为“Y”



Step 4: 计算 Appearance=Good 下的信息增益

$$\mathrm{H(D2)} = -\frac{2}{5}log\frac{2}{5} - \frac{3}{5}log\frac{3}{5} = 0.971$$

$$\text{Income: } \mathrm{H(Inc = Great)} = -\frac{1}{1}log\frac{1}{1} = 0$$

$$\mathrm{H(Inc = Good)} = -\frac{1}{2}log\frac{1}{2} - \frac{1}{2}log\frac{1}{2} = 1$$

$$\mathrm{H(Inc = Low)} = -\frac{2}{2}log\frac{2}{2} = 0$$

$$H(D2|F_{Income}) = \frac{1}{5} * H(Inc = Great) + \frac{2}{5} * H(Inc = Good) + \frac{2}{5} * H(Inc = Low) = 0.4$$

Age: $H(Age = Younger) = -\frac{2}{2} log \frac{2}{2} = 0$

$$H(Age = Older) = -\frac{3}{3} log \frac{3}{3} = 0$$

$$H(D2|F_{Age}) = \frac{2}{5} * H(Age = Younger) + \frac{3}{5} * H(Age = Older) = 0$$

Profession: $H(Prof = Steady) = -\frac{1}{3} log \frac{1}{3} - \frac{2}{3} log \frac{2}{3} = 0.918$

$$H(Prof = Unstable) = -\frac{1}{2} log \frac{1}{2} - \frac{1}{2} log \frac{1}{2} = 1$$

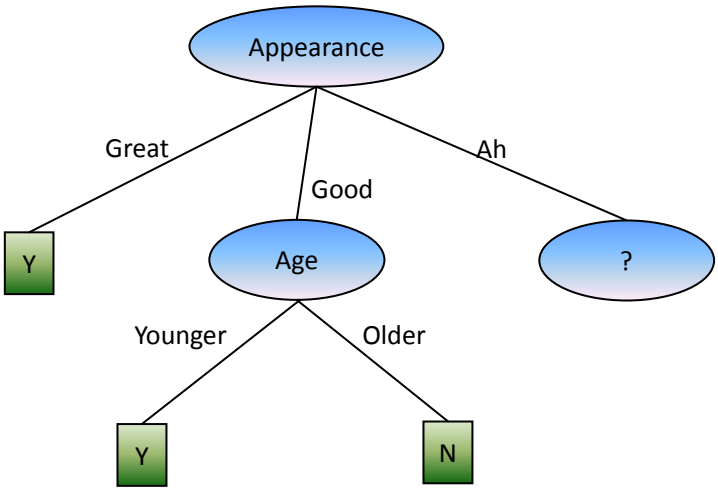
$$H(D2|F_{Prof}) = \frac{3}{5} * H(Prof = Steady) + \frac{2}{5} * H(Prof = Unstable) = 0.951$$

$G(D2|F_{Inc}) = H(D2) - H(D2|F_{Inc}) = 0.971 - 0.4 = 0.371$

$\mathbf{G(D2|F_{Age}) = H(D2) - H(D2|F_{Age}) = 0.971 - 0 = 0.971}$

$G(D2|F_{Prof}) = H(D2) - H(D2|F_{Prof}) = 0.971 - 0.951 = 0.02$

可见，Age 的信息增益最大，取 Age 作为子结点的特征，引出两个节点，一个对应“Younger”，包含 2 个样本，属于同一类，所以是叶节点，类标记为“Y”；另一个对于“N”，包含 3 个样本，属于同一类，所以也是叶节点，类标记为“N”



Step 5: 计算 Appearance=Ah 下的信息增益

$$H(D3) = -\frac{2}{5} log \frac{2}{5} - \frac{3}{5} log \frac{3}{5} = 0.971$$

Income: $H(Inc = Great) = -\frac{1}{2} log \frac{1}{2} - \frac{1}{2} log \frac{1}{2} = 1$

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

$$H(D3|F_{Income}) = \frac{2}{5} * H(Inc = Great) + \frac{3}{5} * H(Inc = Good) = 0.951$$

Age: $H(Age = Younger) = -\frac{2}{3}log\frac{2}{3} - \frac{1}{3}log\frac{1}{3} = 0.918$

$$H(Age = Older) = -\frac{1}{2}log\frac{1}{2} - \frac{1}{2}log\frac{1}{2} = 1$$

$$H(D3|F_{Age}) = \frac{3}{5} * H(Age = Younger) + \frac{2}{5} * H(Age = Older) = 0.951$$

Profession: $H(Prof = Steady) = -\frac{3}{3}log\frac{3}{3} = 0$

$$H(Prof = Unstable) = -\frac{2}{2}log\frac{2}{2} = 0$$

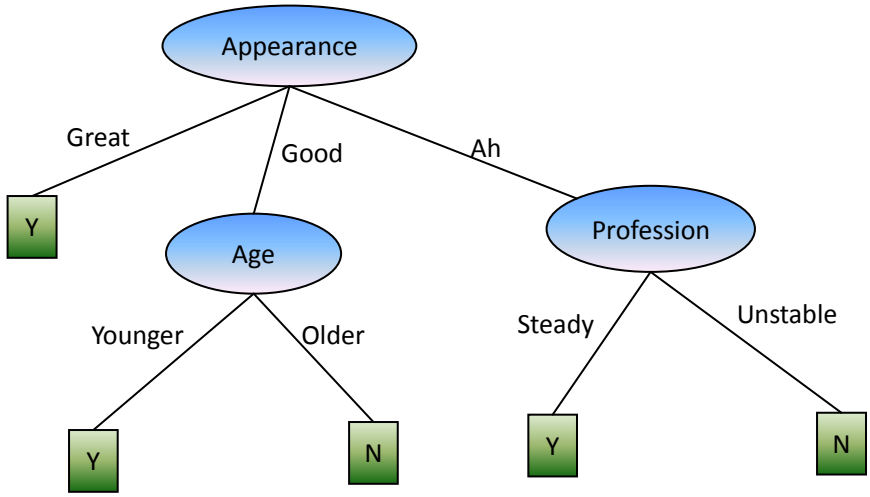
$$H(D3|F_{Prof}) = \frac{3}{5} * H(Prof = Steady) + \frac{2}{5} * H(Prof = Unstable) = 0$$

$$G(D3|F_{Inc}) = H(D3) - H(D3|F_{Inc}) = 0.971 - 0.951 = 0.02$$

$$G(D3|F_{Age}) = H(D3) - H(D3|F_{Age}) = 0.971 - 0.951 = 0.02$$

$$\mathbf{G(D3|F_{Prof}) = H(D3) - H(D3|F_{Prof}) = 0.971 - 0 = 0.971}$$

可见，Profession 的信息增益最大，取 Profession 作为子结点的特征，引出两个节点，一个对应“Steady”，包含 3 个样本，属于同一类，所以是叶节点，类标记为“Y”；另一个对于“N”，包含 2 个样本，属于同一类，所以也是叶节点，类标记为“N”。最终根据 ID3 算法生成的决策树如下所示。



C4.5：根据信息增益率来生成决策树

I. What is Gain Ratio?

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

信息增益率即：

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

II. Why we are prone to use Gain Ratio?

假如每个属性中每种类别都只有一个样本，那这样该属性下每一类的信息熵就等于 0，改属性划分所得到的信息增益最大，但是这种划分没有意义。所以根据信息增益就无法选择出有效分类特征。所以，C4.5 选择使用信息增益率对 ID3 进行改进。

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

以上面的例子说明

$$\text{SplitInfo}(D|F_{App}) = -\frac{4}{14}log\frac{4}{14} - \frac{5}{14}log\frac{5}{14} - \frac{5}{14}log\frac{5}{14} = 1.577$$

$$\text{SplitInfo}(D|F_{Inc}) = -\frac{4}{14}log\frac{4}{14} - \frac{6}{14}log\frac{6}{14} - \frac{4}{14}log\frac{4}{14} = 1.557$$

$$\text{SplitInfo}(D|F_{Age}) = -\frac{7}{14}log\frac{7}{14} - \frac{7}{14}log\frac{7}{14} = 1$$

$$\text{SplitInfo}(D|F_{Prof}) = -\frac{8}{14}log\frac{8}{14} - \frac{6}{14}log\frac{6}{14} = 0.985$$

$$GR(D|F_{App}) = \frac{G(D|F_{App})}{\text{SplitInfo}(D|F_{App})} = \frac{0.246}{1.577} = 0.156$$

$$GR(D|F_{Inc}) = \frac{G(D|F_{Inc})}{\text{SplitInfo}(D|F_{Inc})} = \frac{0.029}{1.557} = 0.019$$

$$GR(D|F_{Age}) = \frac{G(D|F_{Age})}{\text{SplitInfo}(D|F_{Age})} = \frac{0.151}{1} = 0.151$$

$$GR(D|F_{Prof}) = \frac{G(D|F_{Prof})}{\text{SplitInfo}(D|F_{Prof})} = \frac{0.048}{0.985} = 0.049$$

可得，信息增益率最大的为 Appearance，同样是以 Appearance 作为根节点。

CART：使用基尼不纯度（Gini Impurity）来决定划分

IV. What Gini Index?

Gini 指数度量数据划分或训练元组集 D 的不纯度。

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

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

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

取属性 Gini 指数最小的作为划分标准。生成的是二叉树。以上面的数据为例

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=6Y+1N, Older: 7=3Y+4N}

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

Appearance: {Great|Good,Ah} = {4=4Y | 10=5Y+5N}

{Good|Great,Ah} = {5=2Y+3N | 9=7Y+2N}

{Ah|Great,Good} = {5=3Y+2N | 9=6Y+3N}

Appearance: $Gini(\mathrm{D}|\{\mathrm{Great}|\mathrm{Good},\mathrm{Ah}\}) = \frac{4}{14}*\left(1-\left(\frac{4}{4}\right)^2\right)+\frac{10}{14}*\left(1-\left(\frac{5}{10}\right)^2-\left(\frac{5}{10}\right)^2\right) = 0.357$

$Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Ah}\}) = \frac{5}{14}*\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)+\frac{9}{14}*\left(1-\left(\frac{7}{9}\right)^2-\left(\frac{2}{9}\right)^2\right) = 0.394$

$Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Ah}\}) = \frac{5}{14}*\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)+\frac{9}{14}*\left(1-\left(\frac{6}{9}\right)^2-\left(\frac{3}{9}\right)^2\right) = 0.457$

$Gini(\mathrm{D}|\mathrm{F_{App}}) = \mathrm{MIN}(Gini(\mathrm{D}|\{\mathrm{Great}|\mathrm{Good},\mathrm{Ah}\}), Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Ah}\}), Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Ah}\}))$

$= Gini(\mathrm{D}|\{\mathrm{Great}|\mathrm{Good},\mathrm{Ah}) = 0.357$

Income: {Great|Good,Low} = {4=3Y+1N | 10=6Y+4N}

{Good|Great,Low} = {6=4Y+2N | 8=5Y+3N}

{Low|Great,Good} = {4=2Y+2N | 10=7Y+3N}

Income: $Gini(\mathrm{D}|\{\mathrm{Great}|\mathrm{Good},\mathrm{Low}\}) = \frac{4}{14}*\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{10}{14}*\left(1-\left(\frac{6}{10}\right)^2-\left(\frac{4}{10}\right)^2\right) = 0.45$

$Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Low}\}) = \frac{6}{14}*\left(1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2\right)+\frac{8}{14}*\left(1-\left(\frac{5}{8}\right)^2-\left(\frac{3}{8}\right)^2\right) = 0.458$

$Gini(\mathrm{D}|\{\mathrm{Low}|\mathrm{Great},\mathrm{Good}\}) = \frac{4}{14}*\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)+\frac{10}{14}*\left(1-\left(\frac{7}{10}\right)^2-\left(\frac{3}{10}\right)^2\right) = 0.443$

$Gini(\mathrm{D}|\mathrm{F_{Inc}}) = \mathrm{MIN}(Gini(\mathrm{D}|\{\mathrm{Great}|\mathrm{Good},\mathrm{Low}\}), Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Low}), Gini(\mathrm{D}|\{\mathrm{Good}|\mathrm{Great},\mathrm{Low}))$

$= Gini(\mathrm{D}|\{\mathrm{Low}|\mathrm{Great},\mathrm{Good}) = 0.443$

Age: $Gini(\mathrm{D}|\mathrm{Younger}) = 1-\left(\frac{6}{7}\right)^2-\left(\frac{1}{7}\right)^2 = 0.245$

$Gini(\mathrm{D}|\mathrm{Older}) = 1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2 = 0.49$

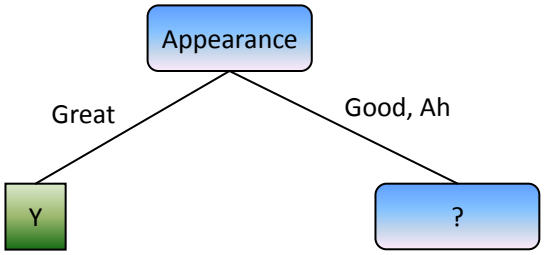
$$Gini(\mathrm{D}|\mathrm{F}_{\mathrm{Age}}) = \frac{7}{14} * Gini(\mathrm{D}|\mathrm{Younger}) + \frac{7}{14} * Gini(\mathrm{D}|\mathrm{Older}) = 0.367$$

Profession: $Gini(\mathrm{D}|\mathrm{Steady}) = 1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{8}\right)^2 = 0.375$

$$Gini(\mathrm{D}|\mathrm{Unstable}) = 1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$$

$$Gini(\mathrm{D}|\mathrm{F}_{\mathrm{Prof}}) = \frac{8}{14} * Gini(\mathrm{D}|\mathrm{Steady}) + \frac{6}{14} * Gini(\mathrm{D}|\mathrm{Unstable}) = 0.429$$

由上可知，Gini 指数最小的属性为 Appearance，以 Great 与(Good, Ah)为根节点分二叉树。



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

1. ID3 只能处理离散数据，不能处理连续数据。
2. ID3 一般会优先选择有较多属性值的 Feature，因为属性值多的 Feature 会有相对较大的信息增益（信息增益反映的给定一个条件以后不确定性减少的程度,必然是分得越细的数据集确定性更高,也就是条件熵越小,信息增益越大）