创建时间:	2019/8/8 15:13	**更新时间:**	2019/8/9 21:45
作者:	Min Xia		
URL:	https://katex.org/docs/supported.html		

ID3是决策树的一种经典的构造算法，内部使用了信息熵和信息增益。 每次迭代选择信息增益最大的特征属性作为分割属性。

信息熵:

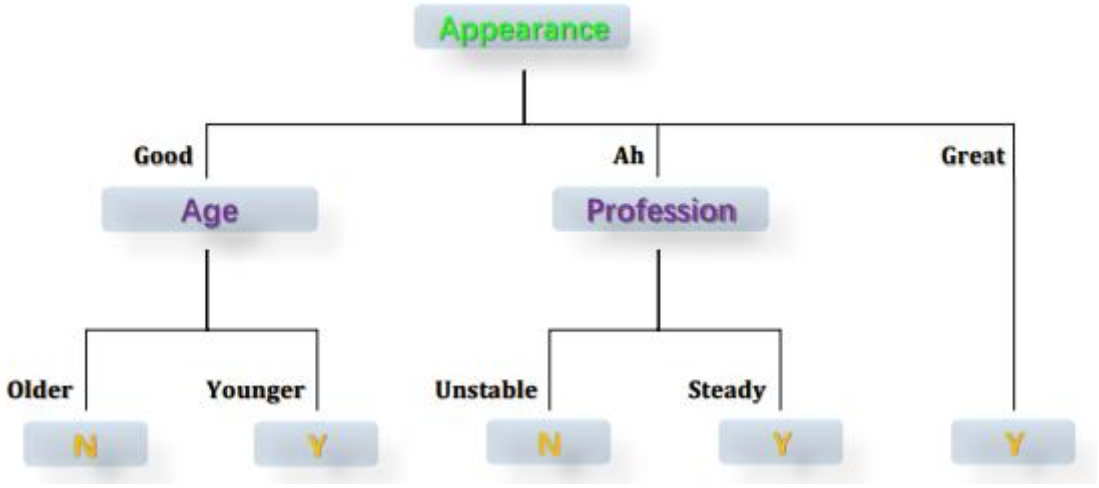
$$Ent(D) = -\sum_{i=1}^{n} P_k \log P_k$$

信息增益:

$$Gain(D, a) = H(D) - \frac{|D^v|}{|D|} \sum_{i=1}^{v} Ent(D^v)$$

example:女婿丈母娘欢迎度

分析：已知总共14人， 9个受欢迎，5个不受欢迎，第一个分割属性为appearance,其中Great下的根节点类别只有一个，无需再分割。



1: Good： 2Y 3N，

Ent(Good) = - 2/5 * log(2/5) - 3/5* log(3/5) = 0.97
从剩余的三个特征(income, age, prefession)里选择分割属性

```
   1: Income:    Low 2N,     Good  1N, 1Y         Great: Y
       H(Income|Good)) = - 1/2 * log(1/2) - 1/2 * log(1/2)  H(Income|Low) =
H(Income|Great)=0
       H(G|income) = Ent(Good) - 2/5 * H{Good)} = 0.57

   2: Age:    Older 3N       Younger 2Y
       H(Age|Older) =  H(Age|Younger) = 0
       H(G|Age) = Ent(Good) - 0 = 0.97

   3: Profession:  steady  2N 1Y  unstable 1N 1Y
       H(Pro|steady) = -2/3 * log(2/3) - 1/3 * log(1/3)=0.756 H(Pro|unsta) =
-1/2 * log(1/2) - 1/2 * log(1/2) = 1
       H(G|prefession) = Ent(Good) - 3/5 * H(Pro|steady)  - 2/5 *
H(Pro|steady) = 0.1164

   信息增益最大的特征属性为Age，此时每个根结点只有一个类别，所以到此分割结束
```

2: Ah: 3Y 2N

Ent(Ah) = - 2/5 * log(2/5) - 3/5* log(3/5) = 0.97
从剩余的三个特征(income, age, prefession)里选择分割属性

```
1: Income:   Good: 2N, 1Y   Great: 1Y 1N
    H(Income|good) = -2/3 * log(2/3) - 1/3 * log(1/3)=0.756  H(Income|great)
= -1/2 * log(1/2) - 1/2 * log(1/2)=1
   H(Ah|Income) = Ent(Ah) - 3/5 * H(Income|good) - 2/5 * H(Income|great)  =
0.1164

2: Age:  Older 1Y 1N   Younger 2Y 1N
   H(Age|older) = -1/2 * log(1/2) - 1/2 * log(1/2)=1  H(Age|younger) = -2/3 *
log(2/3) - 1/3 * log(1/3)
   H(Ah|Age) = Ent(Ah) - 2/5 * (Age|older) -  3/5 * H(Age|younger) = 0.1164

3: profession:  steady: 3Y   Unstable: 2N
   H(Pro|steady) = H(Pro|Unsta) = 0
   H(Ah|Pro) = Ent(Ah) - 0  =  0.97
```

信息增益最大的特征属性为profession，此时每个根结点只有一个类别，所以到此分割结束

3: Question and answer

I. What is Gain Ratio?

$$IV(a) = \sum_{i=1}^{V} \frac{|D^v|}{|D|} \log(\frac{|D^v|}{|D|})$$

$$GainRatio(D) = \frac{Gain(D,a)}{IV(a)}$$

信息增益率：某个特征的信息增益与此特征的分裂信息度量之比。

II. Why we are prone to use Gain Ratio?

信息增益对特征的数目取值较多(更重要是此取值下样本很小或者类别很少了)的偏好，只考虑当前最优分割属性，不考虑全局最优属性；不支持连续值处理，不支持缺失值处理所以对数据适应性相对差；不支持剪枝容易过拟合。

然而信息增益率加入了特征的分裂信息度量，考虑了当前特征与其特征的取值数目之间的关系。

如对于上述问题中选取第一个分割属性时，特征取值数目较多的数值比较大，然而在分母上，所以整体取值会变小，即信息增益率可以修正信息增益的偏好。同时支持连续值处理，支持缺失值处理，支持剪枝。

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

基于信息增益，除以求得的特征的分裂信息度量，信息增益率大的特征作为分割特征。

IV. What Gini Index?

$$Gini(D) = 1 - \sum_{i=1}^{n} p_k(1 - p_k) = 1 - \sum_{i=1}^{n} p_k^2$$

即从数据集D中随机抽取两个样本，其类别标记不一致的概率。所以基尼系数越小，数据集纯度越高.

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

例如：D 为是否欢迎

Gini_index(appearance, Ah) = 1 - (3/5) ** 2 - (2/5) ** 2

Gini_index(appearance, Good) = 1 - (3/5) ** 2 - (2/5)** 2

Gini_index(appearance, Great) =1 - 1

Gini_index(D, appearance) = 5/14 * Gini_index(appearance, Ah)
+ 5/14 * Gini_index(appearance, Good)
+ 4/14 * Gini_index(appearance, Great)

然后对其它特征做同样处理，取基尼系数最小的特征作为分割特征。

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

相比如ID3，ID3对于 特征属性取值数目较多的有所偏好。

CART和C4.5都支持剪枝，不易过拟合、支持缺失值的处理，支持连续值的处理，而且CART可用于分类与回归。