# Name Convention

1. 变量符号及含义

|  |  |
| --- | --- |
|  | 表示 |
| Dataset数据集 | 是samples（ features every sample）  是分类结果 (K classes in total) |
| Node |  |
| Directed Edge |  |

1. 常用词汇

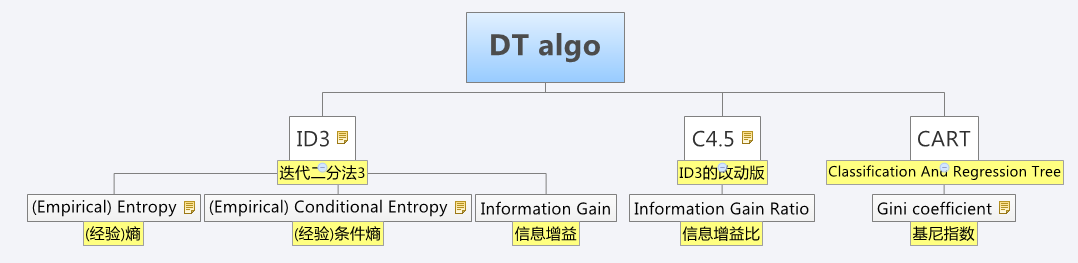
* Decision Tree
* Node
* Internal Node: feature/property
* Leaf node: a class
* Directed edge

# 分类树概况

## 理解分类树

1. 逻辑角度上，decision tree是一组if-then规则的集合，if是feature层面的，then是分类结果层面的。
   1. Root node到leaf node的每一条路径都代表一种if规则，且路径上每一个内部节点都代表一种feature的if条件 # 决策树上的路径或者if规则集合满足特性：互斥且完备， 也就是说每个instance分类结果（then）都有且仅被一条规则（if）覆盖
   2. Leaf node代表then，是instance 分类的结果
2. 概率论上，decision tree是指特定条件下，类的条件概率分布。这一条件概率分布是定义在特征空间（feature space）上的一个划分（partition）， 被划分成一个个互不相交的单元（cell）或区域（region），并在每个cell/region定义概率分布就构成了条件概率分布。

## 分类树思维导图



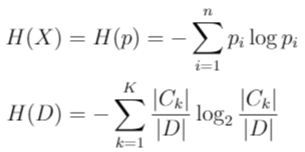
## 分类树步骤

1. Feature selection
2. Decision Tree Generation
3. Decision Tree Pruning

# 分类树算法

## ID3

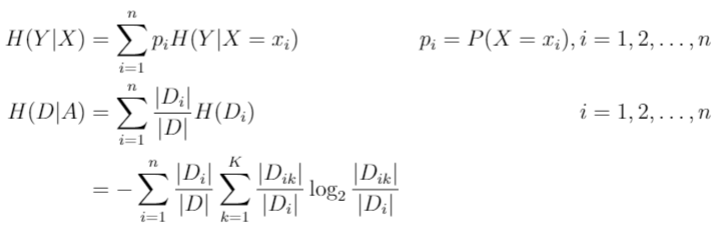
### (Empirical) Entropy (经验)熵



* 熵只与X的分布有关，与X取值无关
* empirical 一次表示 熵中的概率估计是直接用数据分布得到的，而不是原始概率分布函数，有一定的经验性

## (Empirical) Conditional Entropy (经验)条件熵

条件熵$H(Y|X)$表示在已知随机变量$X$的条件下随机变量$Y$的不确定性



## Information Gain信息增益



DT algo

ID3

(Empirical) Entropy

(Empirical) Conditional Entropy

Information Gain

C4.5

Information Gain Ratio

CART

Gini coefficient

C4.5

Information Gain Ratio

CART

Gini coefficient