机器学习数据分析和建模中级教程系列讲座-第3讲

**决策树(Decision Tree)(上)**

主讲人：元俊

**预备知识**

大学物理 初级

数据结构 基础

**上一讲回顾**

LR的究极目标是获得一个稳定的θT向量[w0,w1,w2,w3,…,wn]

LR的窘境?

什么是决策树？

决策树是一个多叉树。

每一个非叶子节点是一个单独维度的判断条件，每一个叶子节点是结论。从根节点开始，经过一次或多次判断得出结论。

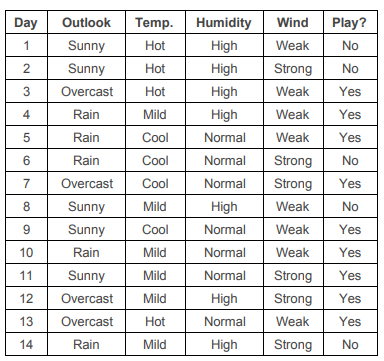
决策树分ID3, C4.5, CART

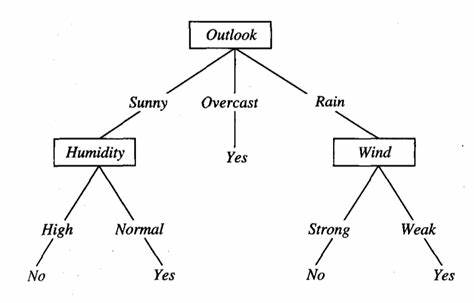
本次课程暂时只涉及二元分类的决策树

决策树的实例

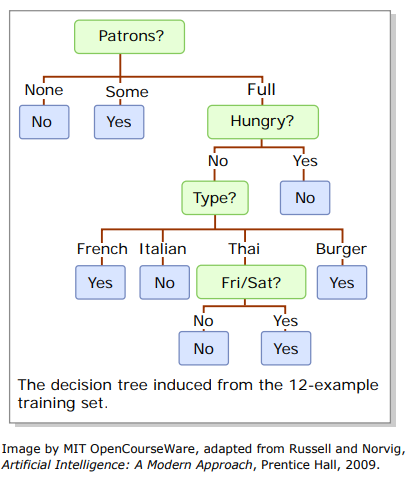
艺考改版—>visio

1.天气和玩耍demo（来自Quinlan的论文）

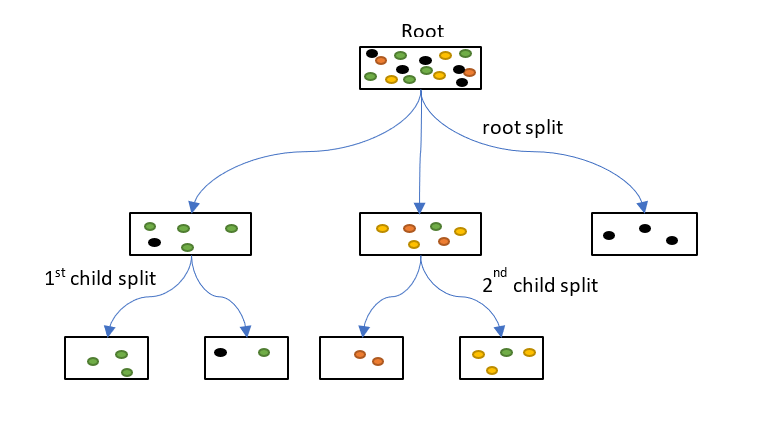




2.饭店Demo（Will the customer wait for a table? MIT ocw）



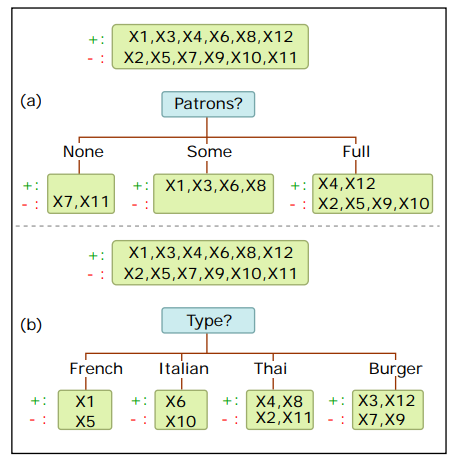
如何建树？



**ID3建树思路**

1. 如果数据集还能分裂，则继续，否则结束
2. 选择一个最佳的特征（feature）和它的分类点，使得这个划分是最有序的（distinctive），如果找不到，则结束
3. 按照特征进行逐次分类，一般来说每一个特征不可重复（ID3，C4.5）
4. 每一个中间节点把数据分裂为若干个不相交的子集
5. 针对每一个子集，递归建树

**如何选特征？？？**



A和B哪个划分方法好？

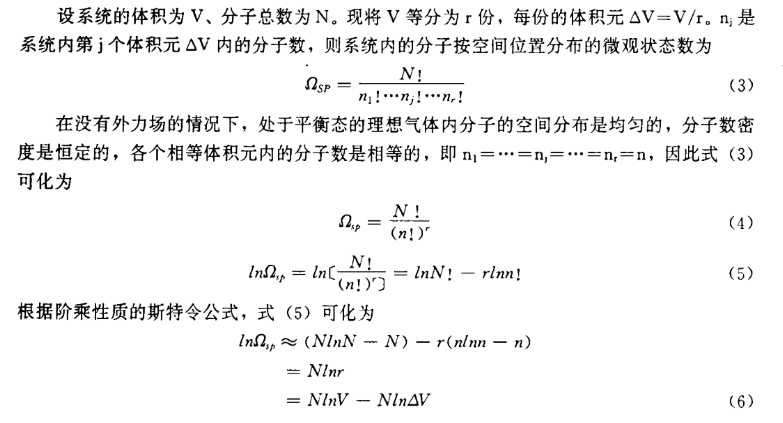
信息熵

混乱和有序，如何量化？



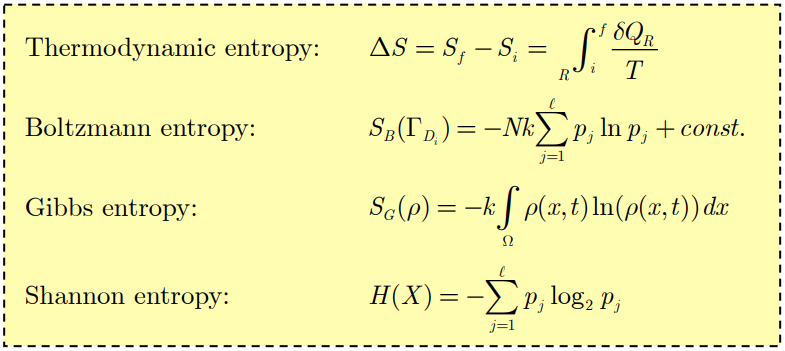
**从热力学到信息学**

Boltzmann熵的理论推导



从Boltzmann到Shannon

Pdf&visio



小知识

1877年，玻尔兹曼用下面的关系式来表示系统无序性的大小：S∝lnΩ。1900年，普朗克引进了比例系数k，将上式写为S=klnΩ。

熵（希腊语：entropia 英语：entropy）的概念是由德国物理学家克劳修斯于1865年所提出。在希腊语源中意为“内在”，即“一个系统内在性质的改变”，公式中一般记为S。

普朗克的签名1923年，德国科学家普朗克（Planck）来中国讲学用到entropy这个词，胡刚复教授翻译时灵机一动，把“商”字加火旁来意译“entropy”这个字，创造了“熵”字（读shāng），因为熵变dS是dQ除以T（温度）的商数。

**熵的范例demo**

**决策树划分的依据--信息增益**

**Summary:**

KEYWORDS: **Decision Tree, Entropy, Shannon Entropy**

决策树非常直观，比较适合低复杂度的枚举类型数据.

熵（Entropy）描述了混沌的量化指标

决策树借用了热力学的熵概念，用信息增益来评估一个划分的量化效果.

Q&A

一些应用场合@MIT

**扩展阅读**

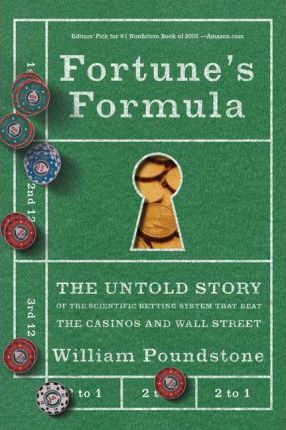
Quilan在1986年的ID3论文<https://hunch.net/~coms-4771/quinlan.pdf>

MIT的决策树在线教程<https://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/lecture-notes/MIT15_097S12_lec08.pdf>

NYU的公开讲义，介绍热力熵，Boltzmann熵，Gibbs熵和Shannon熵4种熵的定义和证明

<http://faculty.poly.edu/~jbain/physinfocomp/lectures/03.BoltzGibbsShannon.pdf>

William Poundstone的*Fortune’s Formula : The Untold Story of the Scientific Betting System That Beat the Casinos and Wall Street*,中文译名财富公式，介绍华尔街和赌场背后的数学概率模型，里面有Claude Shannon的一些传奇故事



**代码附录：**

Git地址：<https://github.com/Yao-lab001/machine_learning_lecture.git>

运行环境：Python 3.7.6 or anaconda 2020.02

预装环境：matplotlib 3.1.3

**下一讲预告**

决策树的具体实现，和各个算法族（ID3,C4.5,CART）的区别

决策树的衍生算法