机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

ML 机器学习

MACHINE LEARNING-MIN ZHANG

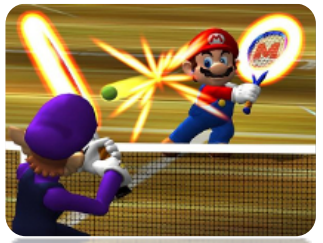
Unit.03

决策树学习

*图片均来自网络或已发表刊物

一、决策树学习基础

一个例子：享受运动



- 已知:

Sky（天气）	Temp（温度）	Humid（湿度）	Wind（风）	Water（水温）	Forecast（预测天气）	Enjoy（享受与否）
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

对于新的一天，他是否可以去享受运动？

适用决策树学习的经典目标问题

- 带有非数值特征的分类问题
- 离散特征
- 没有相似度概念
- 特征无序

另一个例子：水果

- 颜色：红色、绿色、黄色……
- 大小：小、中、大
- 形状：球形、细长
- 味道：甜、酸

样本表示

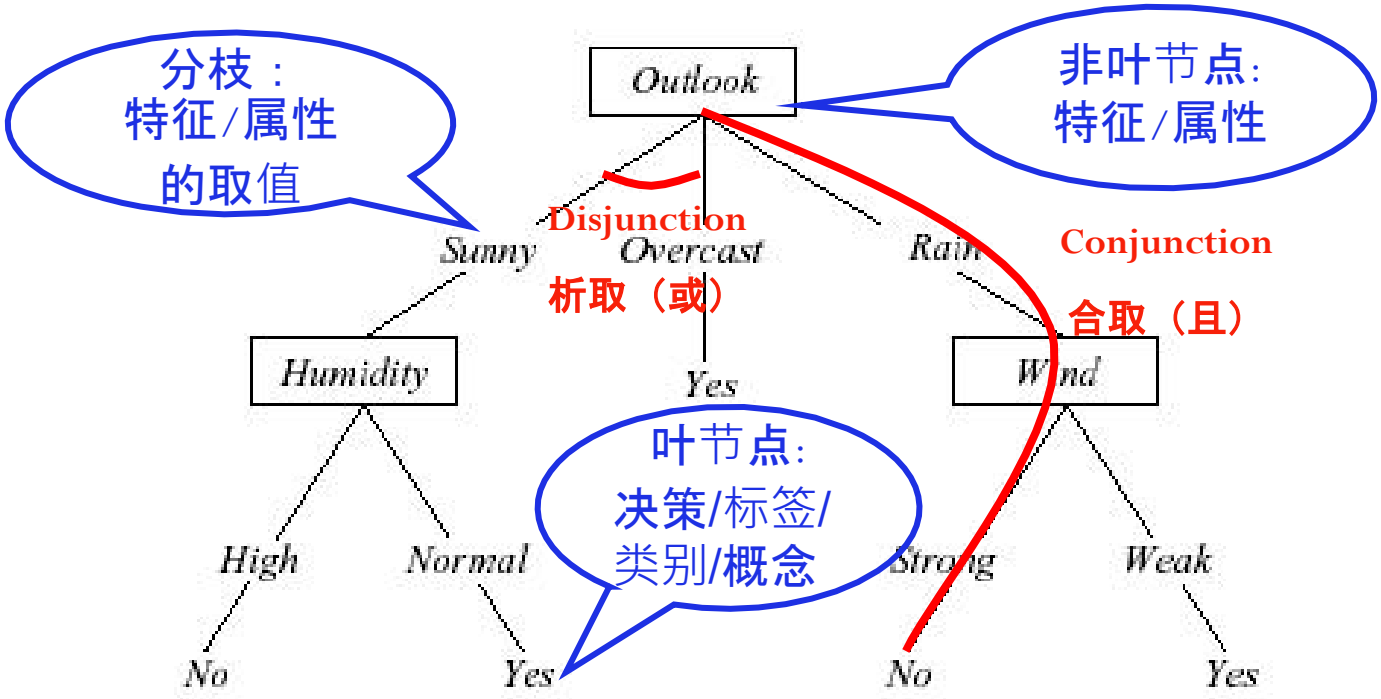
- 属性的列表而非数值向量
- 例如享受运动的例子：
  - 6值属性：*天气*、*温度*、*湿度*、*风*、*水温*、*预测天气*
  - 某一天的天气实例：{晴、暖、一般、强、暖、不变}
- 例如水果的例子：
  - 4值元组：颜色、大小、*形状*、*味道*
  - 某个水果的实例：{红、球形、小、甜}

训练样本

Sky（天气）	Temp（温度）	Humid（湿度）	Wind（风）	Water（水温）	Forecst（预测天气）	Enjoy（享受与否）
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- <香蕉>: 黄色、细长、中、甜
- <西瓜>: 绿色、球形、大、甜
- <香蕉>: 黄色、细长、中、甜
- <葡萄>: 绿色、球形、小、甜
- <葡萄>: 红色、球形、小、酸

决策树 – 概念


分枝：
特征/属性
的取值
非叶节点：
特征/属性
Outlook
Sunny
Overcast
Rain
Disjunction
析取（或）
Conjunction
合取（且）
Humidity
Yes
Wind
叶节点：
决策/标签/
类别/概念
High
Normal
Strong
Weak
No
Yes
No
Yes

决策树发展历史 – 里程碑

- 1966，由Hunt首先提出
- 1970′s~1980′s
  - CART 由Friedman, Breiman提出
  - ID3 由 Quinlan 提出
- 自1990′s以来
  - 对比研究、算法改进(Mingers, Dietterich, Quinlan, etc.)
  - 最广泛使用的决策树算法: C4.5 由 Quinlan 在 1993 年提出

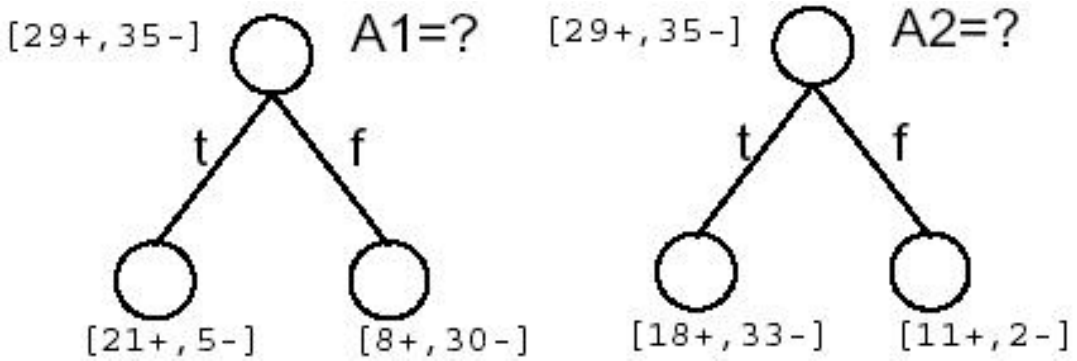
二、经典决策树算法

经典决策树算法 – ID3

Outlook

- 自顶向下，贪心搜索
- 递归算法
- 核心循环:
  - A :下一步 最佳 决策属性
  - 将 A 作为当前节点决策属性
  - 对属性A (v_i)的每个值，创建与其对应的新的子节点
  - 根据属性值将训练样本分配到各个节点
  - 如果 训练样本被完美分类，则退出循环，否则继续下探分裂新的叶节点

ID3 Q1:哪个属性是最佳属性?

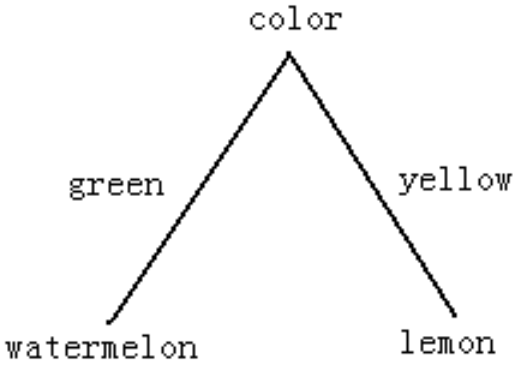


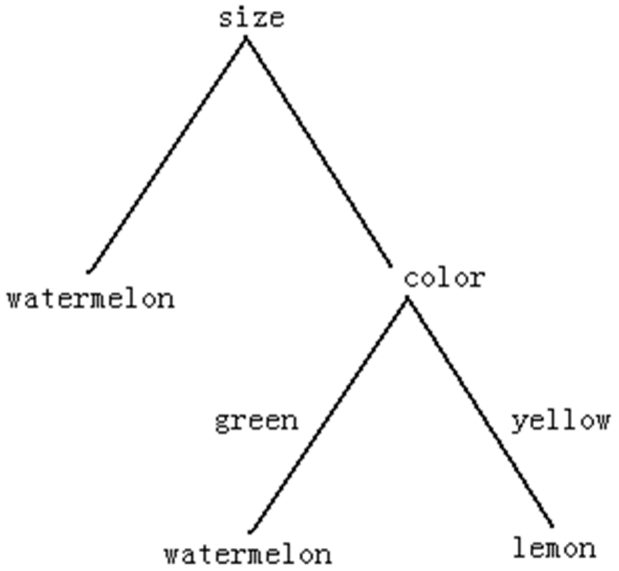
湿度, 风, ... ?

当前最佳属性节点选择

- 基本原则: **简洁**

  —— 我们偏向于使用简洁的具有较少节点的树





属性选择和节点混杂度（Impurity）

- 基本原则：**简洁**

—— 我们偏向于使用简洁的具有较少节点的树

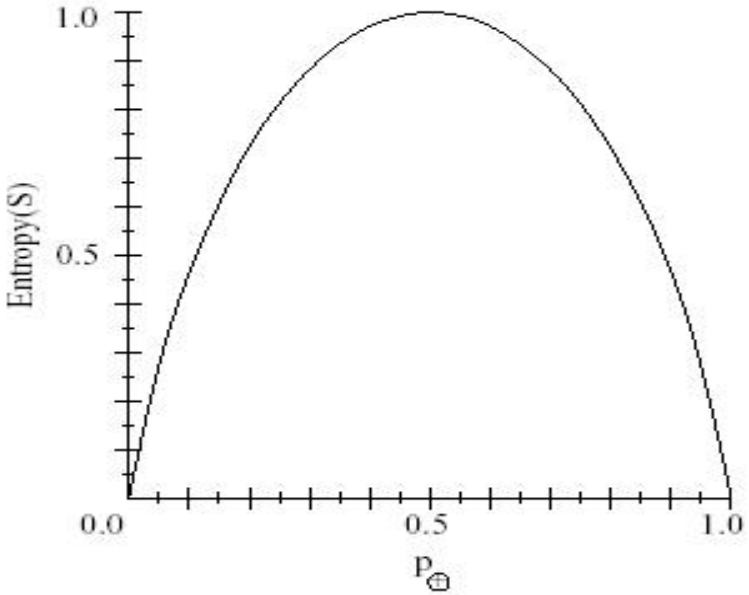
- 在每个节点 N上，我们选择一个属性 T，使得到达当前派生节点的数据尽可能 “纯”
- 纯度(purity) – 混杂度(impurity)

如何衡量混杂度?

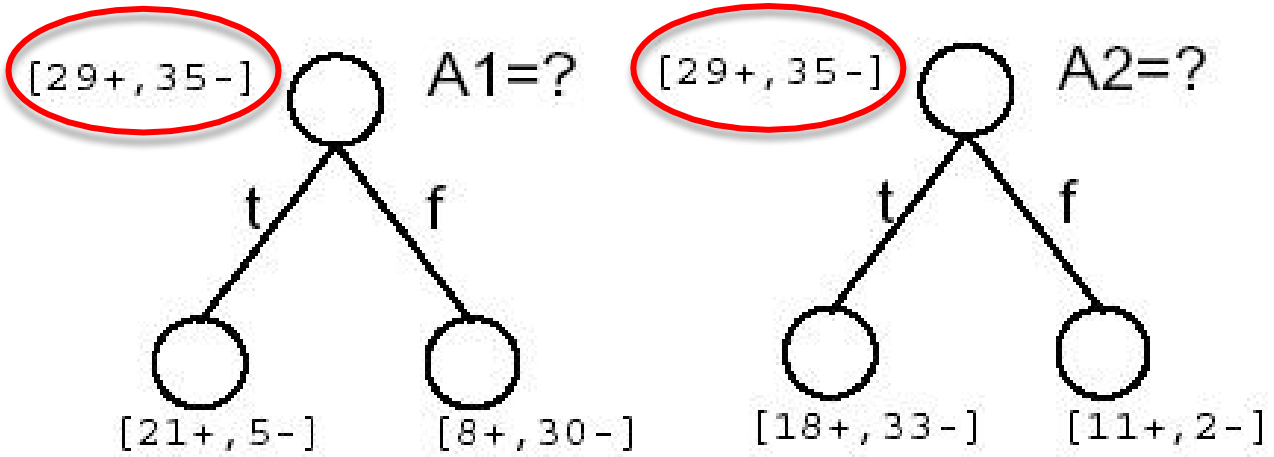
熵(Entropy) (广泛使用)

$$Entropy(N) = -\sum_j P(w_j)\log_2 P(w_j)$$

- 定义: 0log0=0
- 在信息理论中，熵度量了信息的**纯度/混杂度**，或者信息的不确定性
- 正态分布 – 具有最大的熵值



熵



$$Entropy(S) = -\frac{29}{64} \times \log_2 \frac{29}{64} - \frac{35}{64} \times \log_2 \frac{35}{64} = 0.993$$

除了熵以外的其他混杂度度量

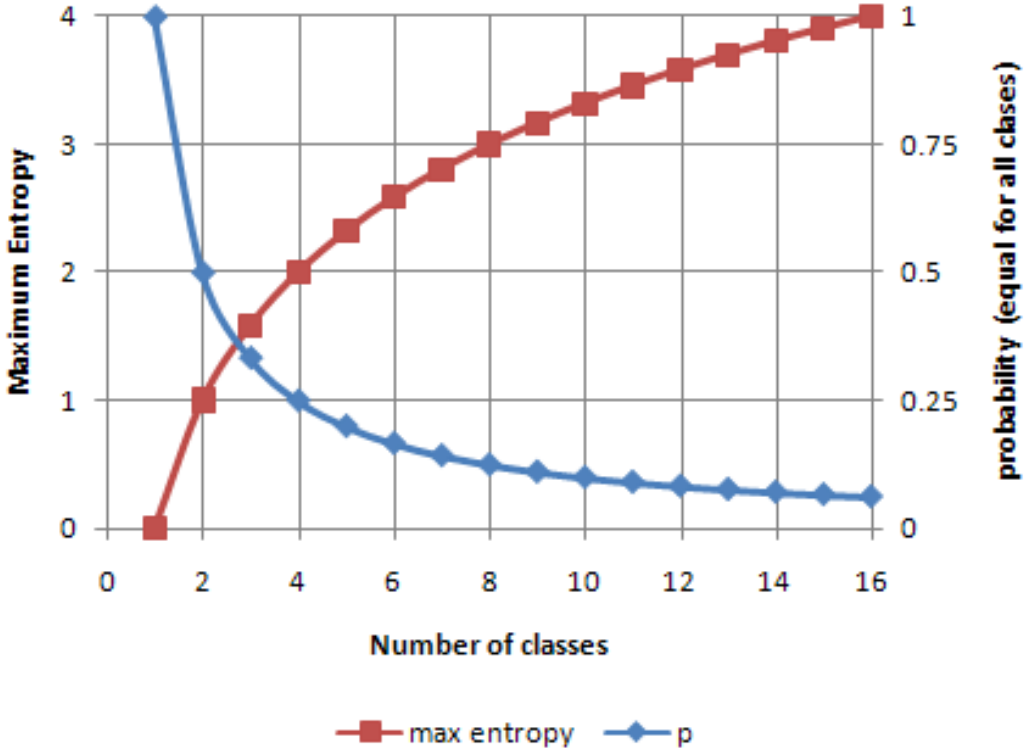
- Gini 混杂度 (Duda 倾向于使用 Gini 混杂度)

$$i(N)=\sum_{i\neq j}P(w_i)P(w_j)=1-\sum_j P^2(w_j)$$

- 错分类混杂度

$$i(N)=1-\max_j P(w_j)$$

混杂度 (熵)

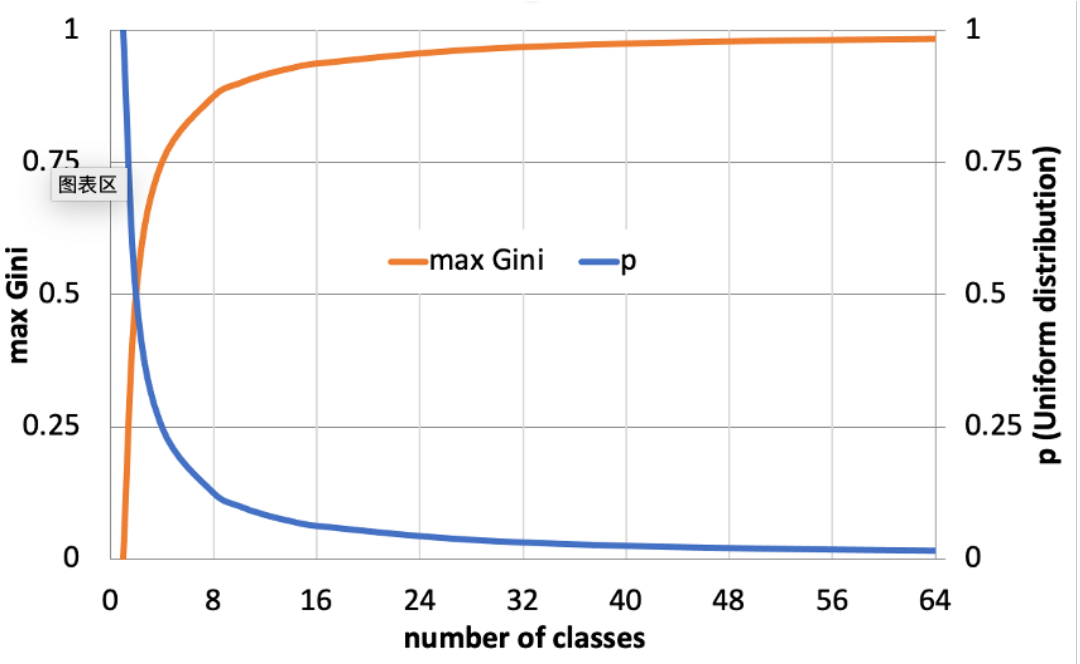


$$Entropy(N) = -\sum_i P(w_j)\log_2 P(w_j)$$

混杂度 (Gini)

$$i(N)=\sum_{i\neq j}P(w_i)P(w_j)=1-\sum_j P^2(w_j)$$

最大 Gini 混杂度 在 1-n*(1/n)^2=1-1/n 时取得

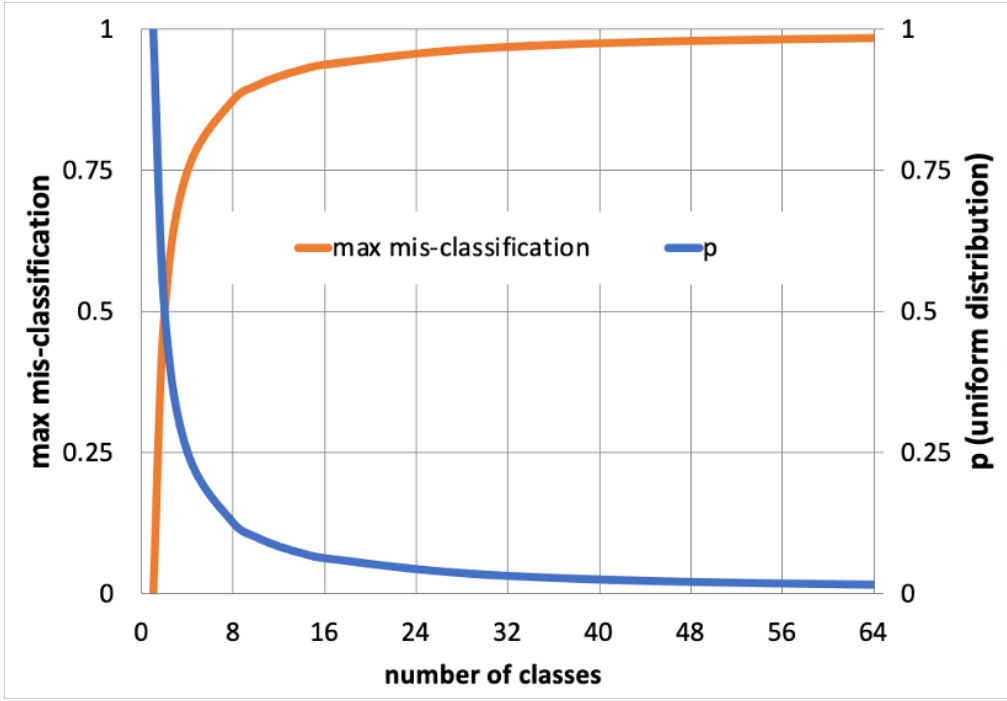


混杂度 (错分类)

$$i(N) = 1 - \max_j P(w_j)$$

最大 Gini 混杂度 在 1-n*(1/n)^2=1-1/n 时取得

在有n类时，最大错分类混杂度 = 最大Gini混杂度 1-max{1/n}=1-1/n



度量混杂度的变化ΔI(N) — 信息增益(IG)，例如

- 由于对A的排序整理带来的熵的期望减少量

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

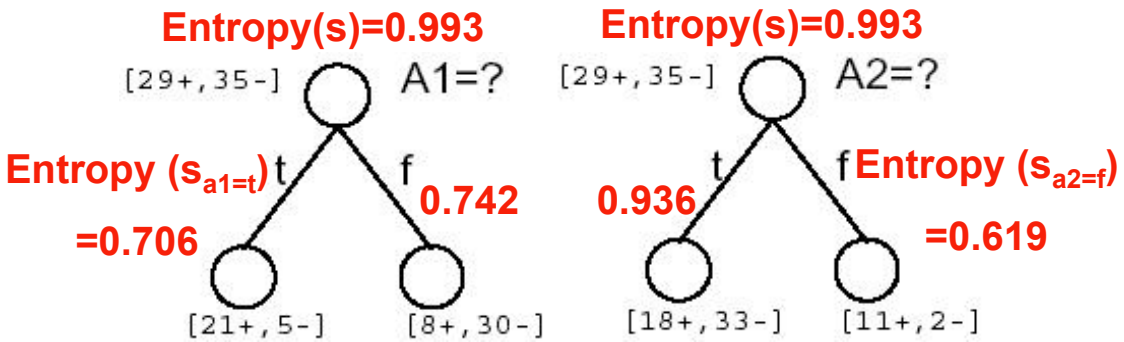
原始S的熵值

经过属性A分类以后的期望熵值

信息增益, IG

- 由于对A的排序整理带来的熵的期望减少量

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

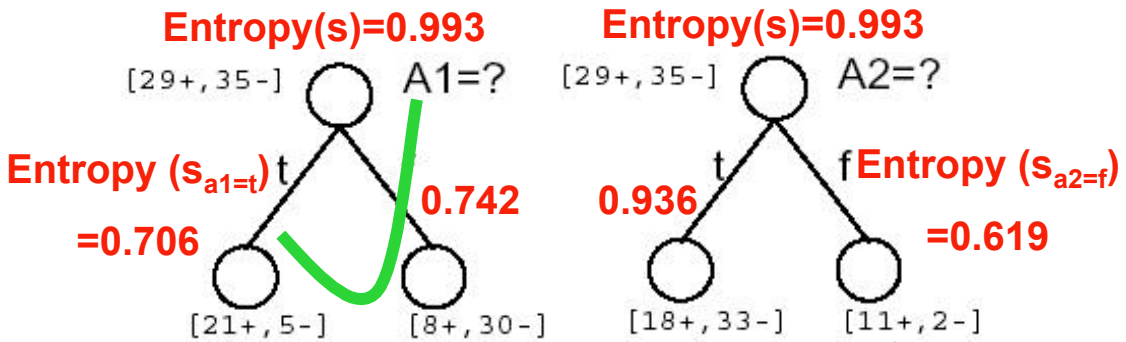


$$Entropy(S) = -\frac{29}{64} \times \log_2 \frac{29}{64} - \frac{35}{64} \times \log_2 \frac{35}{64} = 0.993$$

信息增益, IG

- 由于对A的排序整理带来的熵的期望减少量

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$



$$Gain(S,A_1) = 0.993 - (\frac{26}{64} \times 0.706 + \frac{38}{64} \times 0.742) = 0.266, \quad Gain(S,A_2) = 0.121$$

ID3 Q2: 何时返回(停止分裂节点)？

- “如果训练样本被完美分类”
- 情形 1: 如果当前子集中所有数据 有完全相同的输出类别，那么终止
- 情形 2: 如果当前子集中所有数据 有完全相同的输入特征，那么终止

可能的情形3: 如果 所有属性分裂的信息增益为0, 那么终止

这是个好想法吗？

ID3 Q2 :何时返回(停止分裂节点)

- y=a XOR b

a	b	y
0	0	0
0	1	1
1	0	1
1	1	0

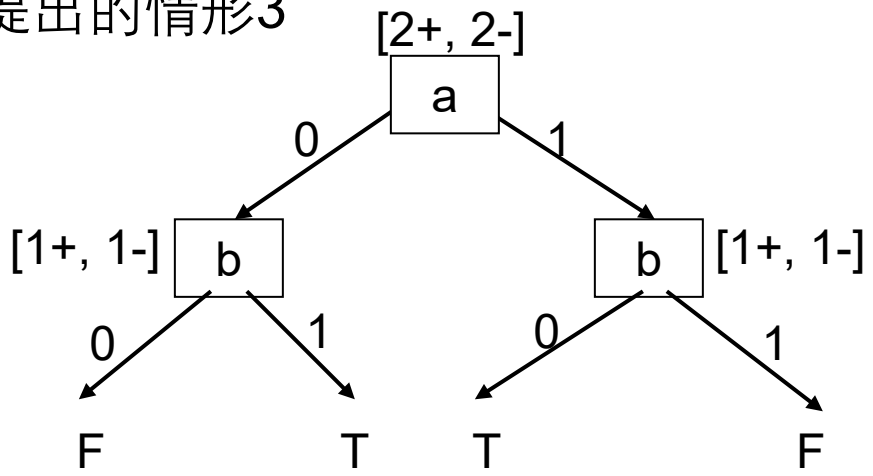
信息增益:

属性	值	概率分布	IG
a	0	50%	0
	1	50%	
b	0	50%	0
	1	50%	

- 根据提出的情形3，甚至在第一步就无法选择任何属性

ID3 Q2：何时返回？

- 如果不考虑提出的情形3



在ID3中，只有2种情形会停止分裂：

- 相同的输出类别 或 相同的输入特征值

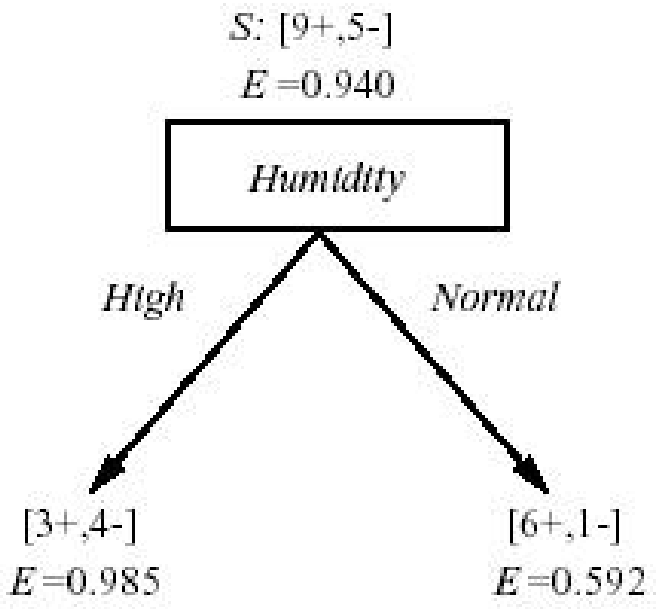
讨论：如果它们有相同的输入特征值但不同的输出类别，意味着什么？

ID3 举例：训练样本

High: 3+,4 -; Normal: 6+,1- Total: 9+, 5- ;

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

ID3 举例: 选择特征



Gain(S,Humidity)

=0.940 – [(7/14)*0.985 + (7/14)*0.592]

=0.151

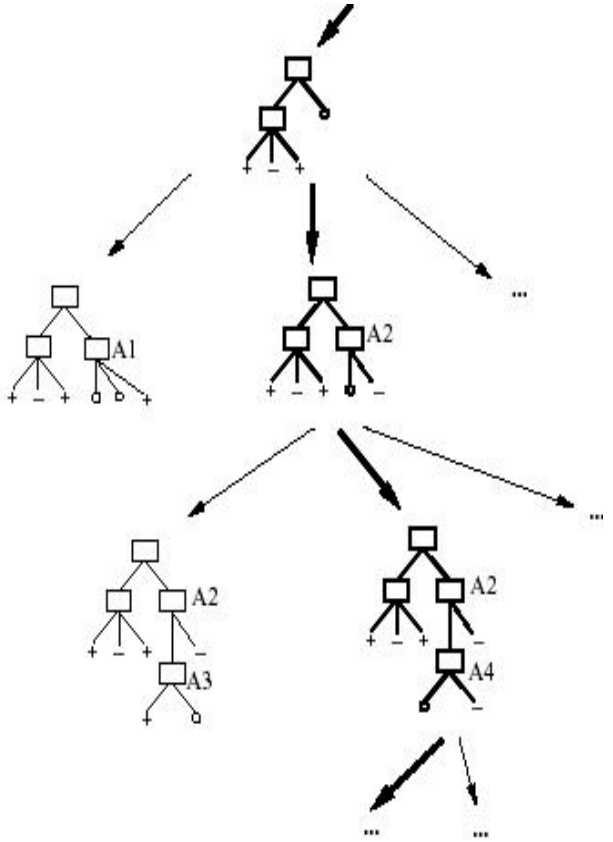
Gain(S,Outlook)=0.246

Gain(S,Wind)=0.048

Gain(S,Temperature)=0.029

ID3算法搜索的假设空间

- 假设空间是完备的
  - 目标函数一定在假设空间里
- 输出单个假设
  - 不超过20个问题(根据经验)
- 没有回溯
  - 局部最优
- 在每一步中使用子集的所有数据
  - 数据驱动的搜索选择
  - 对噪声数据有鲁棒性



ID3中的归纳偏置 （Inductive Bias）

- 假设空间 H 是作用在样本集合 X 上的
  - 没有对假设空间作限制
- 偏向于在靠近根节点处的属性具有更大信息增益的树
  - 尝试找到最短的树
  - 该算法的偏置在于对某些假设具有一些偏好 (搜索偏置)， 而不是对假设空间 H 做限制(描述偏置).
  - 奥卡姆剃刀 （Occam's Razor） *：偏向于符合数据的最短的假设

*仅介绍想法，不作更细致讨论。更多信息可参考:

Domingos, The role of Occam' s Razor in knowledge discovery. Journal of Data Mining and Knowledge Discovery, 3(4), 1999.

CART (分类和回归树)

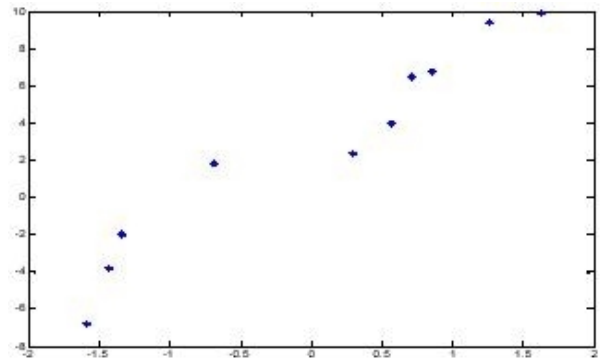
一个通用的框架:

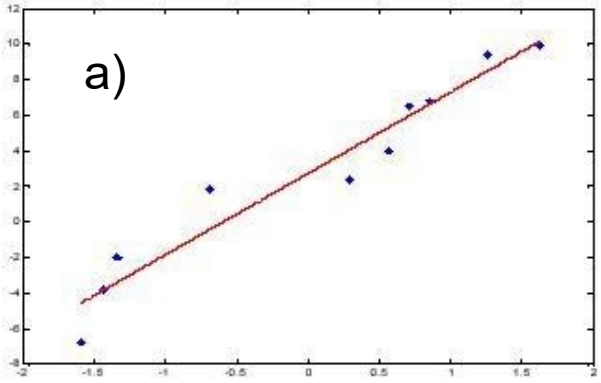
- 根据训练数据构建一棵决策树
- 决策树会逐渐把训练集合分成越来越小的子集
- 当子集纯净后不再分裂
- 或者接受一个不完美的决策

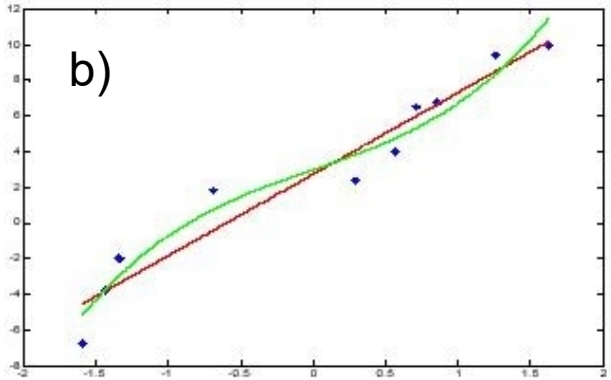
许多决策树算法都在这个框架下，包括ID3、C4.5等等。

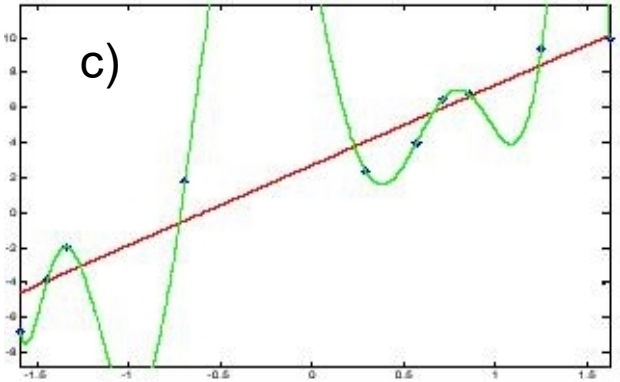
三、过拟合问题

什么是过拟合？

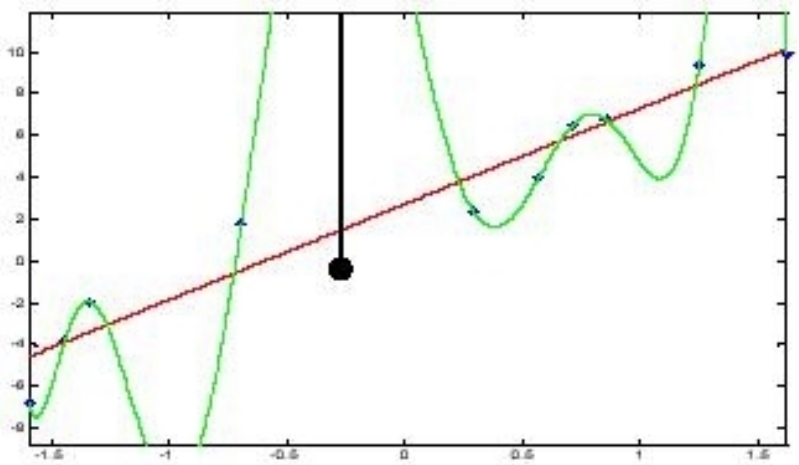








什么是过拟合？



- 我们说$h \in H$ 对训练集过拟合，如果存在另一个假设 $h' \in H$ 满足:

$$err_{train}(h) < err_{train}(h')$$

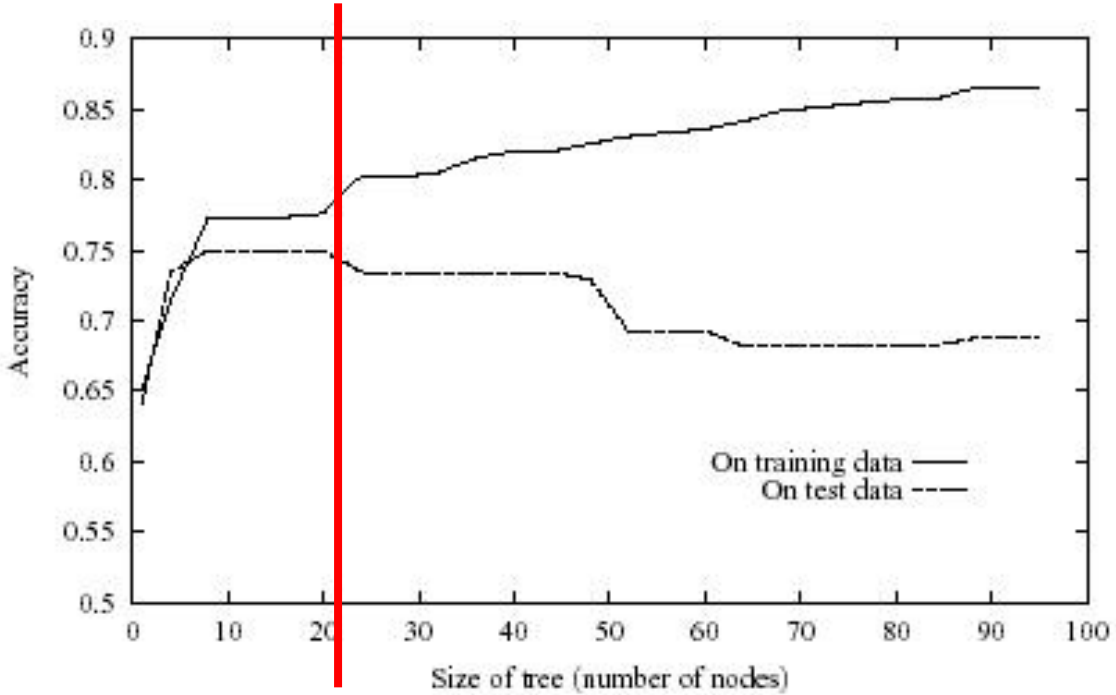
AND

$$err_{test}(h) > err_{test}(h')$$

决策树过拟合的一个极端例子：

- 每个叶节点都对应单个训练样本 —— 每个训练样本都被完美地分类
- 整个树相当于仅仅是一个数据查表算法的简单实现

决策树学习中的过拟合


0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
Accuracy
0
10
20
30
40
50
60
70
80
90
100
Size of tree (number of nodes)
On training data
On test data

3.1 决策树总结

- 介绍及基本概念
- 以ID3算法为例
  - 算法描述
  - 选择特征
  - 终止条件
  - ID3算法的归纳偏置
- 过拟合问题