


第1.2节 决策树学习 (Decision Tree)



内容

- ✓ 决策树的基本原理和算法
- ✓ 熵、信息增益和特征选择
- ✓ 决策树学习中的过拟合问题
- ✓ 交叉验证与树的修剪

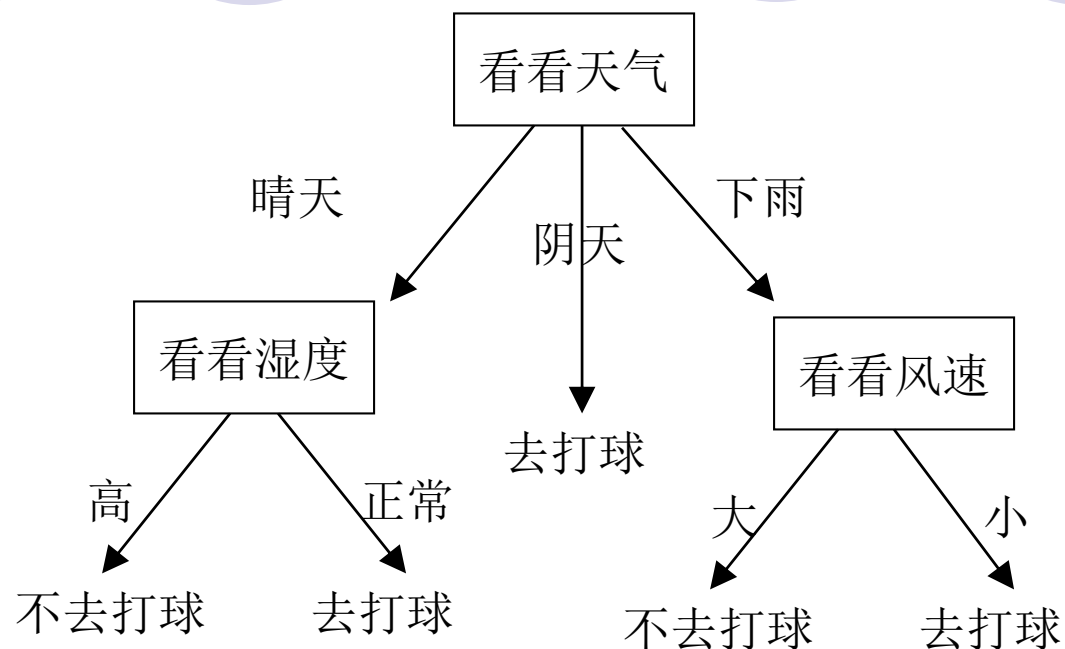
如何根据下表数据学习一个是否去打球的模型？

编号	天气	温度	湿度	风	是否去打球
1	晴天	炎热	高	弱	不去
2	晴天	炎热	高	强	不去
3	阴天	炎热	高	弱	去
4	下雨	适中	高	弱	去
5	下雨	寒冷	正常	弱	去
6	下雨	寒冷	正常	强	不去
7	阴天	寒冷	正常	强	去
8	晴天	适中	高	弱	不去
9	晴天	寒冷	正常	弱	去
10	下雨	适中	正常	弱	去
11	晴天	适中	正常	强	去
12	阴天	适中	高	强	去
13	阴天	炎热	正常	弱	去
14	下雨	适中	高	强	不去

内容

- ✓ 决策树的基本原理和算法
- ✓ 熵、信息增益和特征选择
- ✓ 决策树学习中的过拟合问题
- ✓ 交叉验证与树的修剪

决策树学习——决定是否去打球



节点：每一个节点测试一维特征, x_i

分支：特征的可选数值（此处为离散值）

叶子节点：最终预测 Y or $P(Y | Y \in Leaf)$

基本的决策树学习算法—(ID3)

***node* = root**

循环

{

1. 为当下一个节点选择一个最好的属性 **x**
2. 将属性**x**分配给节点***node***
3. 对于**x**的所有可能数值，创建一个降序排列的节点***node***
4. 将所有训练样本在叶子节点排序分类
5. 如果分类结果达到了错误率要求，跳出循环，否则，
在叶子节点开始新循环-> 递归

}

基本的决策树学习算法—(ID3)

- ID3的思想

- 自顶向下构造决策树
- 从“哪一个特征将在树的根节点被测试”开始
- 使用统计测试来确定每一个实例特征单独分类训练样例的能力

- ID3的过程

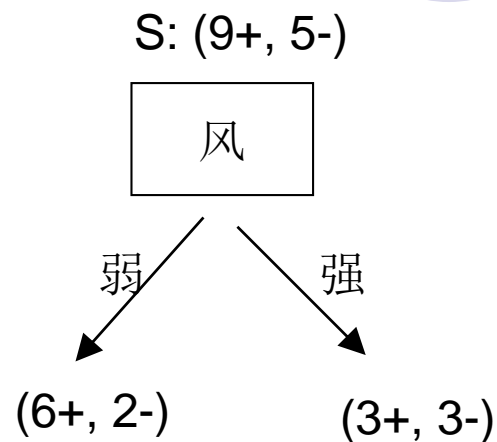
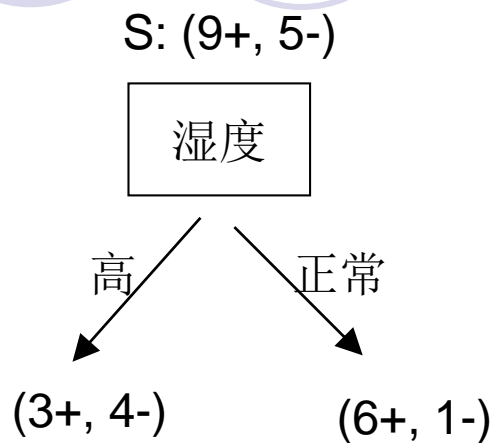
- 分类能力最好的特征被选作树的根节点
- 根节点的每个可能值产生一个分支
- 训练样例排列到适当的分支
- 重复上面的过程

基本的决策树学习算法——(ID3)

表-1：是否去打球的数据统计——训练数据

编号	天气	温度	湿度	风	是否去打球
1	晴天	炎热	高	弱	不去
2	晴天	炎热	高	强	不去
3	阴天	炎热	高	弱	去
4	下雨	适中	高	弱	去
5	下雨	寒冷	正常	弱	去
6	下雨	寒冷	正常	强	不去
7	阴天	寒冷	正常	强	去
8	晴天	适中	高	弱	不去
9	晴天	寒冷	正常	弱	去
10	下雨	适中	正常	弱	去
11	晴天	适中	正常	强	去
12	阴天	适中	高	强	去
13	阴天	炎热	正常	弱	去
14	下雨	适中	高	强	不去

决策树学习原理简介—(ID3)



问题：哪一个属性（特征）更好？

内容

- ✓ 决策树的基本原理和算法
- ✓ 熵、信息增益和特征选择
- ✓ 决策树学习中的过拟合问题
- ✓ 交叉验证与树的修剪

熵

熵:物理学概念

宏观上：热力学定律一体系的熵变等于可逆过程吸收或耗散的热量除以它的绝对温度（克劳修斯，1865）

微观上：熵是大量微观粒子的位置和速度的分布概率的函数，是描述系统中大量微观粒子的无序性的宏观参数（波尔兹曼，1872）

结论：熵是描述事物无序性的参数，熵越大则无序性越强,在信息领域定义为“熵越大，不确定性越大”（香浓，1948年）



熵

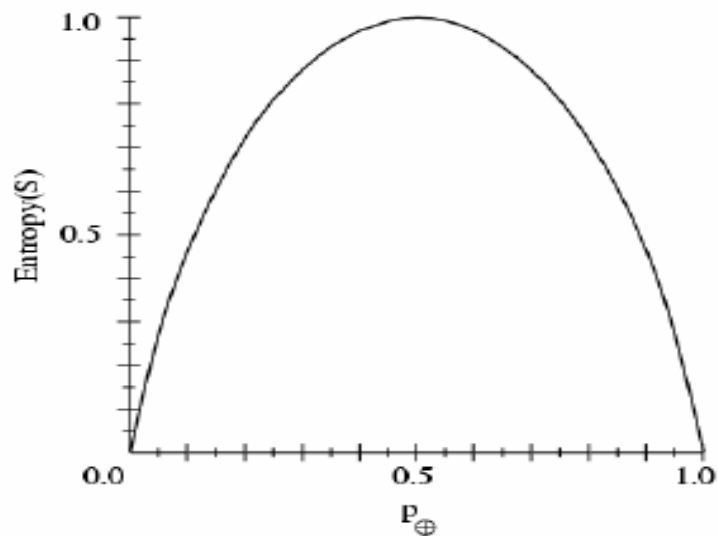
随机变量的熵 $I(X)$

$$I(X) = -\sum_{i=1}^n P(X=i) \log_2 P(X=i)$$

熵 比较多的用于信源编码，数据压缩，假设是最有效的编码方式是使用 位编码 $X=i$
于是对于随即变量的最有效编码位之和：

$$\sum_{i=1}^n P(X=i) (-\log_2 P(X=i))$$

熵



S 表示训练集中的样本

p_{\ominus} 表示训练集中反例样本的比例

p_{\oplus} 表示训练集中正例样本的比例

$I(S) = -p_{\oplus} \log_2(p_{\oplus}) - p_{\ominus} \log_2(p_{\ominus})$ 表示训练集的熵

信息增益(Information Gain)

- ✓ 信息的增加意味着不确定性的减少，也就是熵的减小；
- ✓ 信息增益在诸多系统中定义为：
 - ✓ 在某一个操作之前的系统熵与操作之后的系统熵的差值
 - ✓ 也即是不确定性的减小量

信息增益(Information Gain)

- 选择特征的标准：选择具有最大信息增益(Information Gain)的特征
- 假设有两个类, + 和 -
 - 假设集合**S**中含有**p**个类别为+的样本,**n**个类别为-的样本
 - 将**S**中已知样本进行分类所需要的期望信息定义为:

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

信息增益(Information Gain)

- 假设特征 x 将把集合 S 划分成 K 份 $\{S_1, S_2, \dots, S_K\}$
 - 如果 S_i 中包含 p_i 个类别为 “+” 的样本, n_i 个类别为 “-”, 的样本。那么划分后的熵就是:

$$E(x) = \sum_{i=1}^K \frac{p_i + n_i}{p + n} I(p_i, n_i)$$

- 在 x 上进行决策分枝所获得的信息增益为:

$$Gain(x) = I(p, n) - E(x)$$

信息增益(Information Gain)

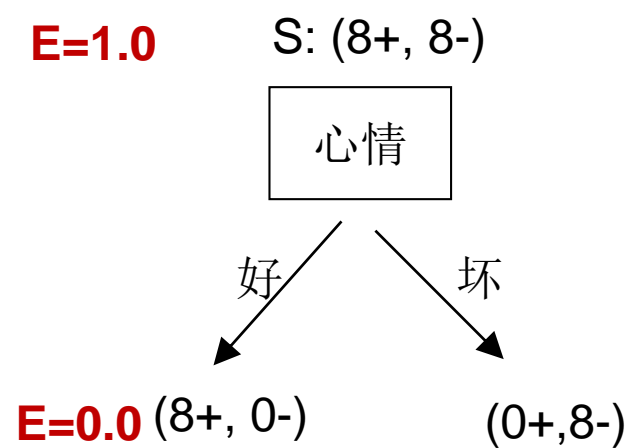
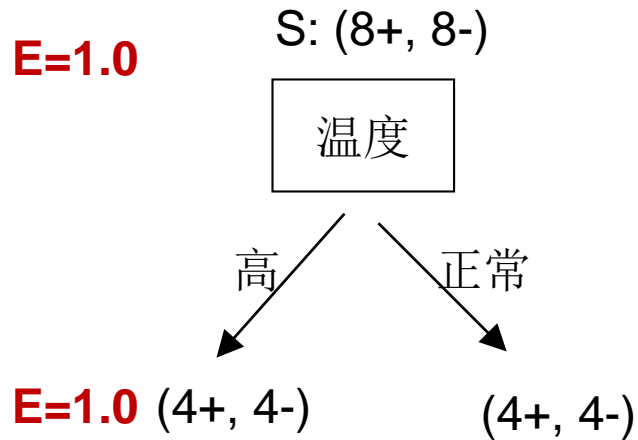
$Gain(S, x)$ 表示给定特征 x 后不确定性的减少，即信息增益
表示了特征与数据集合的互信息

$$Gain(S, x) = Entropy(S) - \sum_{v \in Values(x)} \left| \frac{S_v}{S} \right| Entropy(S_v)$$

$x = v$ 的子集

信息增益(Information Gain)

问题：哪一个属性（特征）更好？分析极端的情况



$$\begin{aligned}\text{Gain}(S, \text{温度}) \\ &= 1.0 - (8/16) * 1.0 - (8/16) * 1.0 \\ &= 0.0\end{aligned}$$

$$\begin{aligned}\text{Gain}(S, \text{心情}) \\ &= 1.0 - (8/16) * 0.0 - (8/16) * 0.0 \\ &= 1.0\end{aligned}$$

下表中湿度和风哪个特征增益更大？

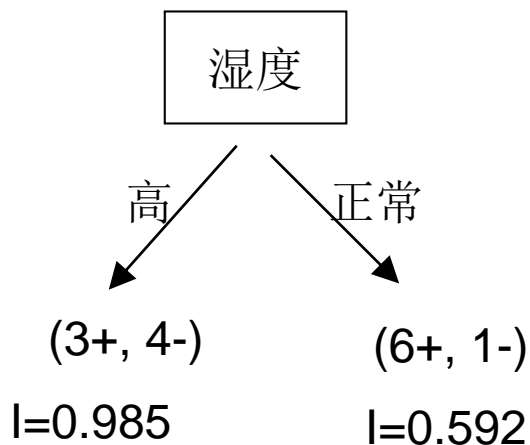
编号	天气	温度	湿度	风	是否去打球
1	晴天	炎热	高	弱	不去
2	晴天	炎热	高	强	不去
3	阴天	炎热	高	弱	去
4	下雨	适中	高	弱	去
5	下雨	寒冷	正常	弱	去
6	下雨	寒冷	正常	强	不去
7	阴天	寒冷	正常	强	去
8	晴天	适中	高	弱	不去
9	晴天	寒冷	正常	弱	去
10	下雨	适中	正常	弱	去
11	晴天	适中	正常	强	去
12	阴天	适中	高	强	去
13	阴天	炎热	正常	弱	去
14	下雨	适中	高	强	不去

信息增益(Information Gain)

问题：哪一个属性（特征）更好？

E=0.940

S: (9+, 5-)



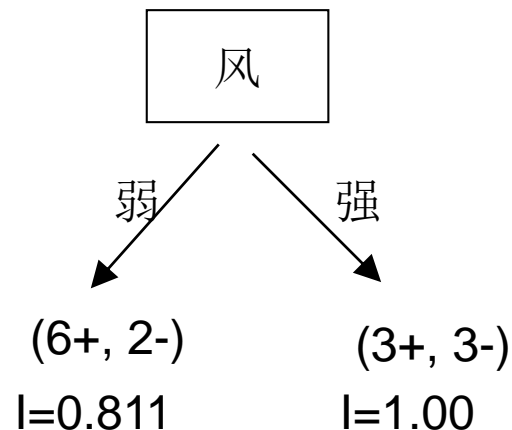
Gain(S, 湿度)

$$=0.940 - (7/14) \cdot 0.985 - 7/14 \cdot 0.592$$

$$=0.151$$

E=0.940

S: (9+, 5-)



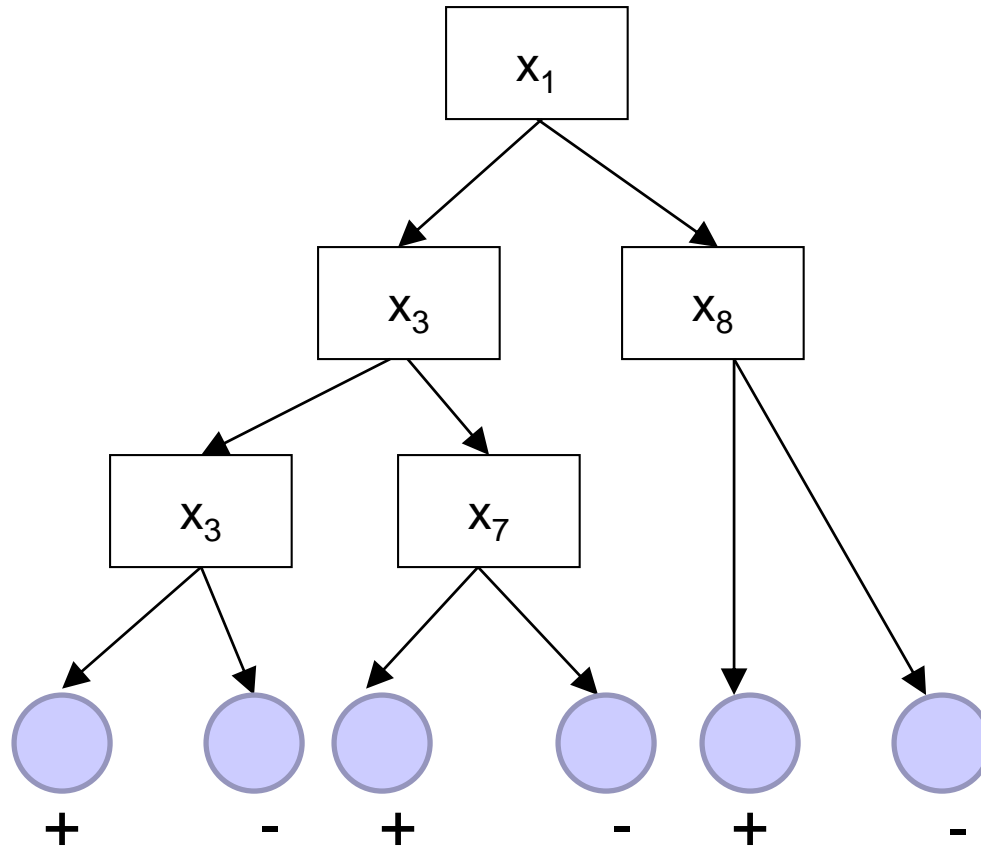
Gain(S, 风)

$$=0.940 - (8/14) \cdot 0.811 - (6/14) \cdot 1.0$$

$$=0.048$$

基本的决策树学习算法—(ID3)

决策树的构造过程示意



基本的决策树学习算法—模型

将树转化为规则

- 将树转化为规则集合
- 测试规则是否相互矛盾
- 将规则排序存储
- Tree:
 - If(阴天) -> 去打球
 - If (晴天)
 - If (风速低) then 去打球
 - Else 不去打球
 -

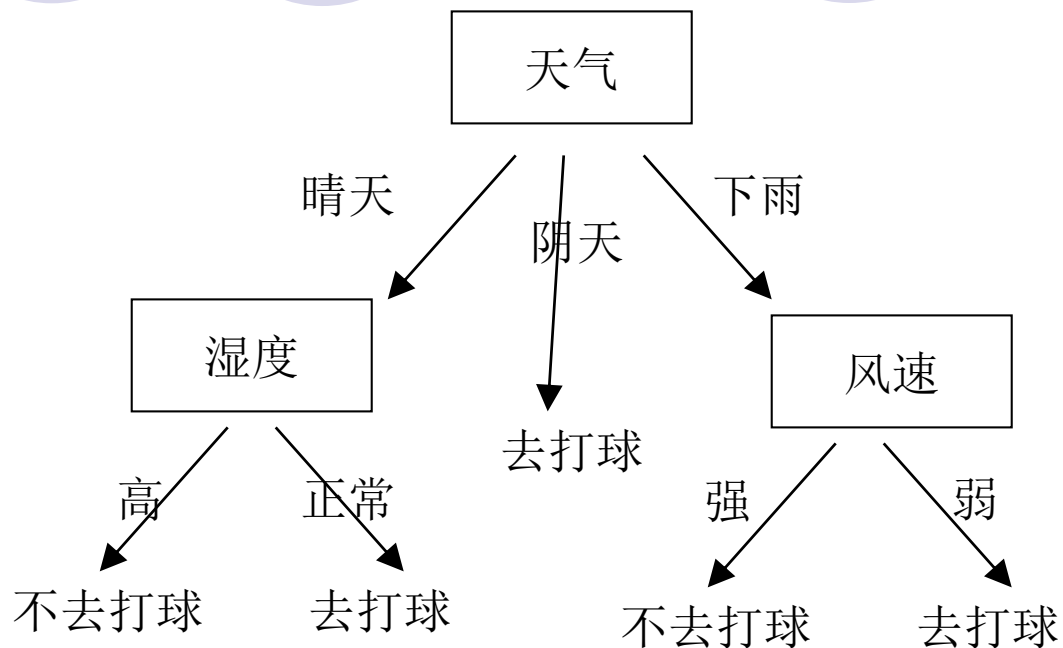
内容

- ✓ 决策树的基本原理和算法
- ✓ 熵、信息增益和特征选择
- ✓ 决策树学习中的过拟合问题
- ✓ 交叉验证与树的修剪

下表中湿度和风哪个特征增益更大？

编号	天气	温度	湿度	风	是否去打球
1	晴天	炎热	高	弱	不去
2	晴天	炎热	高	强	不去
3	阴天	炎热	高	弱	去
4	下雨	适中	高	弱	去
5	下雨	寒冷	正常	弱	去
6	下雨	寒冷	正常	强	不去
7	阴天	寒冷	正常	强	去
8	晴天	适中	高	弱	不去
9	晴天	寒冷	正常	弱	去
10	下雨	适中	正常	弱	去
11	晴天	适中	正常	强	去
12	阴天	适中	高	强	去
13	阴天	炎热	正常	弱	去
14	下雨	适中	高	强	不去

决策树学习的over-fitting



测试样本	晴天	炎热	高	强	去打球
------	----	----	---	---	-----

在使用学习到的决策树模型时，**此测试样本出现错误**

决策树学习的over-fitting

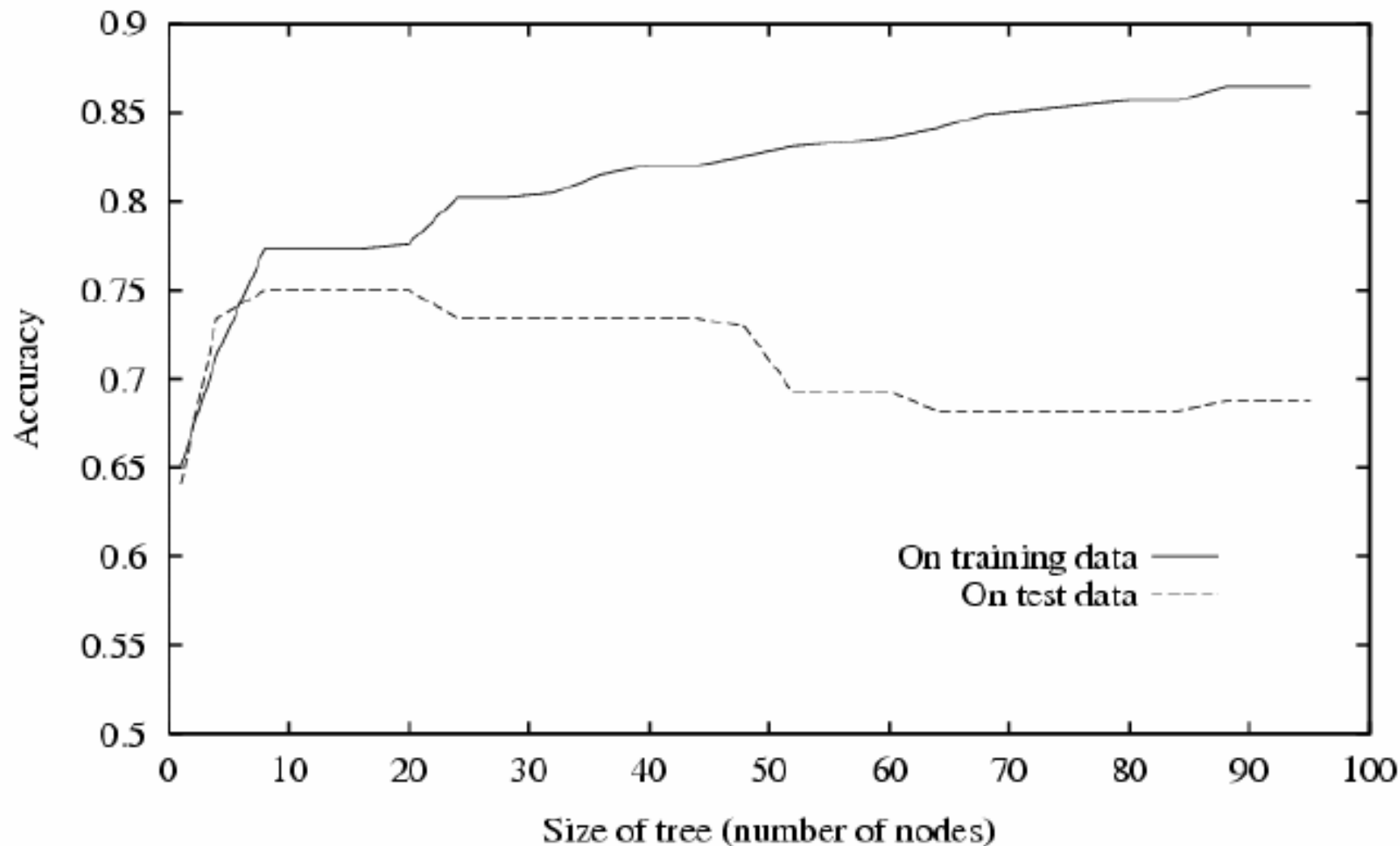
- Over fitting:过拟合

- 对于一个模型假设，当存在其他的假设对**训练样本集合**的拟合比它差，但事实上在**整个样本集合**上上表现得却更好时，我们说这个假设过度拟合训练样例
- **定义：** 给定一个假设空间 H ， $h \in H$ ，如果存在其他的假设 $h' \in H$ ，使得在**训练样本集合**上 h 的错误率比 h' 小，但在**整个样本集合**上， h' 的错误率比 h 小，那么就说假设 h 过度拟合训练数据。

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

$$error_D(h) < error_D(h')$$

决策树学习的over-fitting



决策树学习的over-fitting

- 导致过度拟合的原因

- 一种可能原因是训练样例含有随机错误或噪声
- 当训练数据没有噪声时，过度拟合也有可能发生，特别是当少量的样例被关联到叶子节点时，很可能出现巧合的规律性，使得一些特征恰巧可以很好地分割样例，但却与实际的目标函数并无关系。

决策树学习及over-fitting

避免过拟合的方法

- 如果对数据划分没有明显好处的属性不选择，同时不再将决策数细分
- 构建完成整个树以后进行剪枝
- 在训练数据上测量性能
- 在交叉验证数据上测量性能
- MDL

Minimize

(Size(tree)+Size(misclassifications(tree))

内容

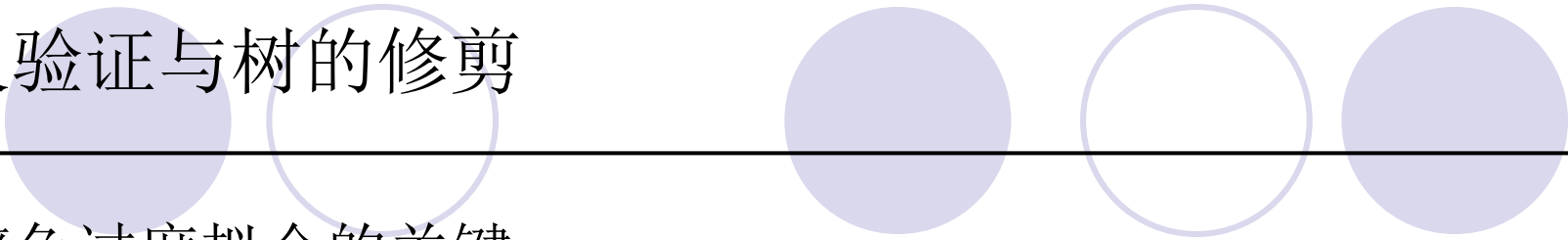
- ✓ 决策树的基本原理和算法
- ✓ 熵、信息增益和特征选择
- ✓ 决策树学习中的过拟合问题
- ✓ 交叉验证与树的修剪

交叉验证与树的修剪

A decorative graphic at the top of the slide features a horizontal black line. Above this line, there are four circles. The first, third, and fourth circles from the left are filled with a light purple color. The second circle is an outline with a light purple border and is not filled. The circles are positioned such that they appear to be behind the horizontal line.

- 避免过度拟合的方法
 - 及早停止树增长
 - 树的修剪
- 两种方法的特点
 - 第一种方法更直观
 - 第一种方法中，精确地估计何时停止树增长很困难
 - 第二种方法被证明在实践中更成功

交叉验证与树的修剪



- 避免过度拟合的关键
 - 使用什么样的准则来确定最终正确树的规模
- 解决方法
 - 使用与训练样例截然不同的一套分离的样例，来评估通过后修剪方法从树上修建节点的效用。
 - 使用所有可用数据进行训练，但进行统计测试来估计扩展（或修剪）一个特定的节点是否有可能改善在训练集合外的实例上的性能。
 - 使用一个明确的标准来衡量训练样例和决策树的复杂度，当这个编码的长度最小时停止树增长。

交叉验证与树的修剪

● 方法评述

- 第一种方法是最普通的，常被称为交叉验证法。
- 可用数据分成两个样例集合：
 - 训练集合，形成学习到的假设
 - 验证集合，评估这个假设在后续数据上的精度
- 方法的动机：即使学习器可能会被训练集合误导，但验证集合不大可能表现出同样的随机波动
- 验证集合应该足够大，以便它本身可提供具有统计意义的实例样本。
- 常见的做法是，样例的三分之二作训练集合，三分之一作验证集合。

交叉验证与树的修剪

- 将树上的每一个节点作为修剪候选对象
- 修剪步骤
 - 删除以此节点为根的子树，使它成为叶结点
 - 把和该节点关联的训练样例的最常见分类赋给它
 - 反复修剪节点，每次总是选取那些删除后可以最大提高决策树在验证集合上的精度的节点
- 继续修剪，直到进一步的修剪是有害的为止
- 数据分成多个子集
 - 训练样例，形成决策树
 - 验证样例，修剪决策树
 - 测试样例，精度的无偏估计

交叉验证与树的修剪

- 从训练集合推导出决策树，增长决策树直到尽可能好地拟合训练数据，允许过度拟合发生
- 将决策树转化为等价的规则集合，方法是为从根节点到叶节点的每一条路径创建一条规则
- 通过“任何能导致估计精度提高的前提”来修剪每一条规则
- 按照修剪过的规则的估计精度对它们进行排序，并按这样的顺序应用这些规则来分类后来的实例

本章作业

本章作业：写出“利用决策树建立转基因植物生物安全评价”
读书报告”

格式为 PPT或者Word，素材见课程网站