



大数据导论

Introduction to Big Data



大数据分类：基础概念与算法

叶允明

计算机科学与技术学院

哈尔滨工业大学（深圳）

目录

- 分类的基本概念
- 决策树分类方法
- 集成学习分类方法

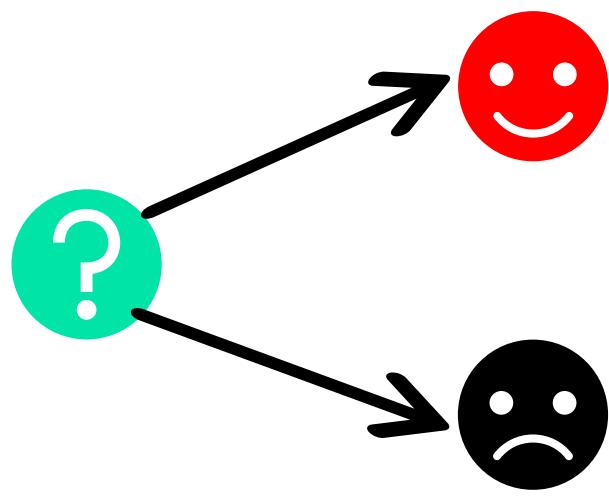
主要参考资料

- Jiawei Han, Micheline Kamber, Jian Pei著；范明，孟小峰等译. 数据挖掘：概念与技术. 机械工业出版社, ISBN: 9787111391401, 2012.
- 第8章

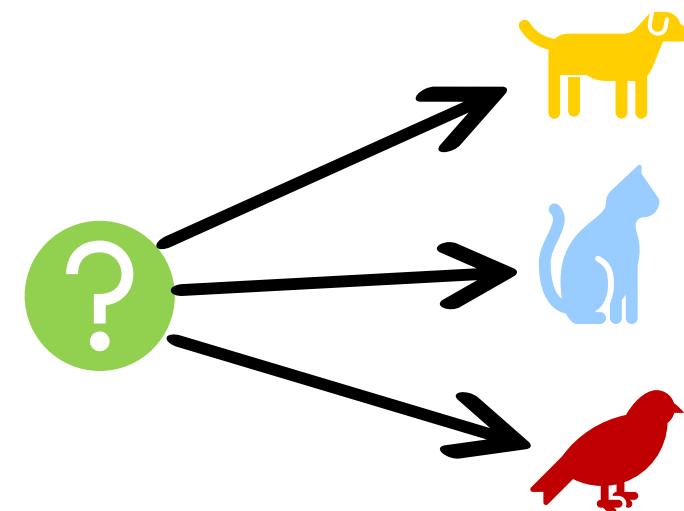
分类的基本概念

分类的基本概念

- 分类 (classification) : 判断 (预测) 给定数据对象所属的类别



二分类



多分类

分类的应用领域

- 分类是人类认识世界的最基本方法
- 几乎每个人工智能应用领域都涉及到分类问题
 - 信用评估
 - 图像识别
 - 目标市场营销
 - 医学诊断
 - 欺诈检测
 - 文本分类
 -

分类任务的定义

- 分类任务可以用一个形式化函数表示：

$$y = f(\textcolor{red}{x}),$$

其中 $\textcolor{red}{x} \in D, y \in \{c_1, c_2, \dots, c_k\}$

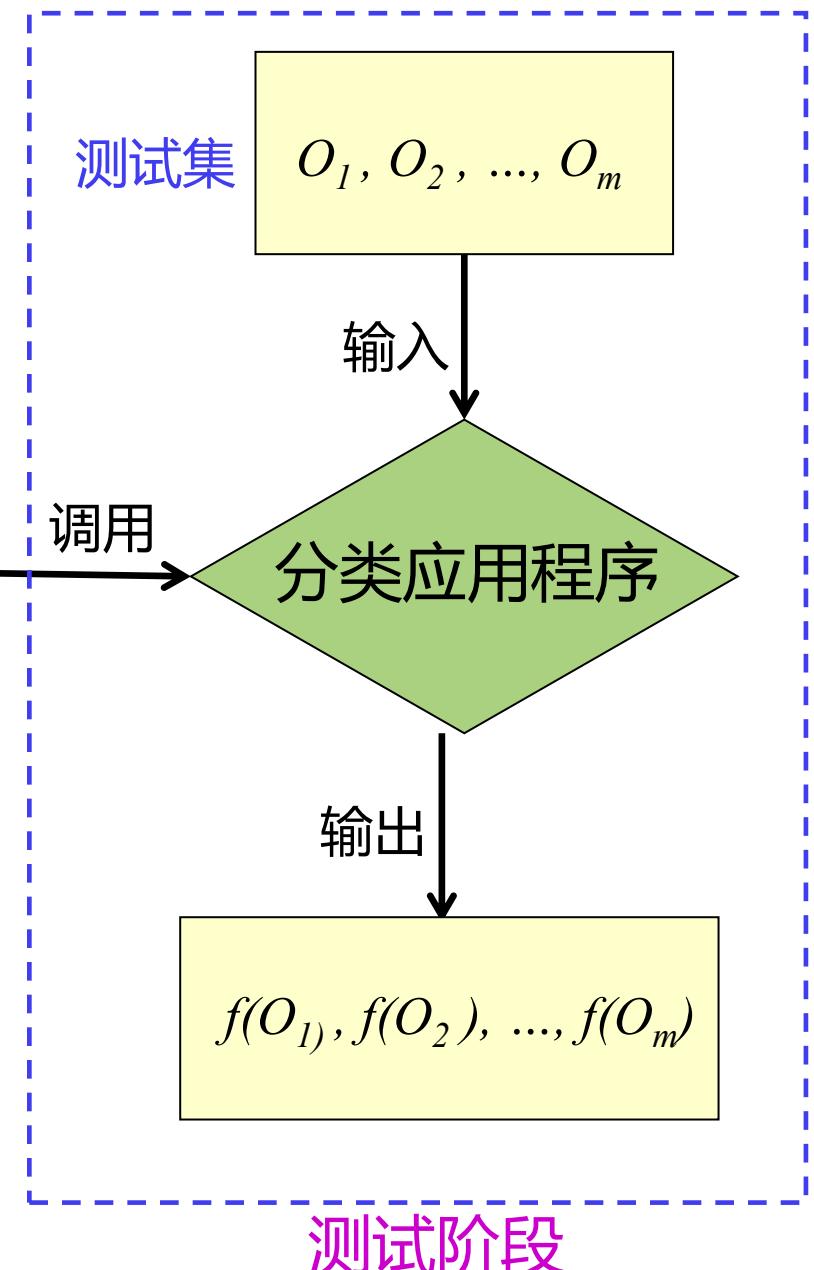
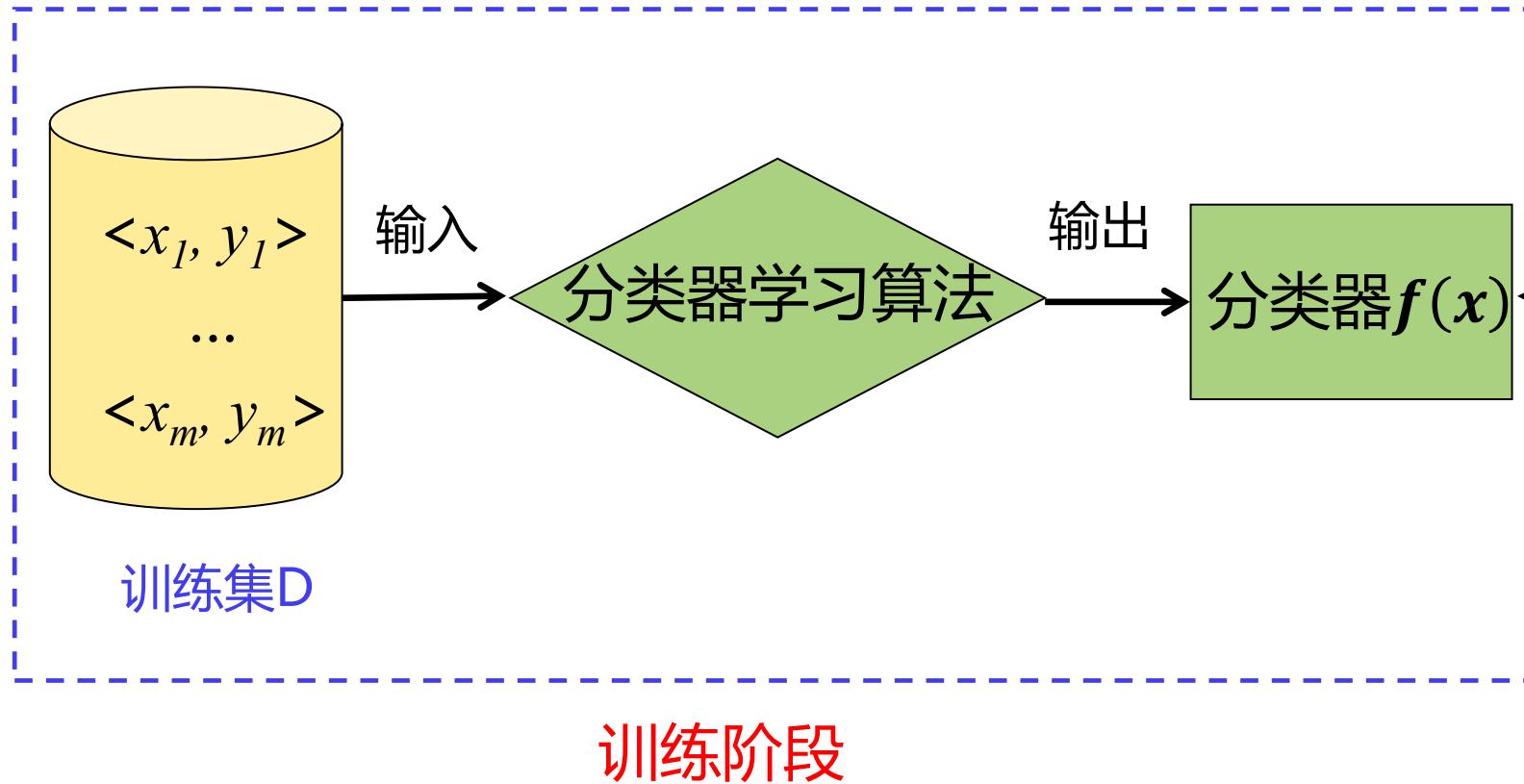
- 分类函数 $f(\textcolor{red}{x})$ 经过运算可以输出一个离散值 y , 又称为“分类器” (classifier)
- 给定数据集 D 中的一个数据对象向量 $\textcolor{red}{x}$ (称为 “实例”)
- y 的取值范围是类别的数字编码集合 (c_1, c_2, \dots, c_k)

如何构造函数 $f(\textcolor{red}{x})$ 呢?

完成分类任务的“两阶段”流程

- 分类器构建（训练阶段）：即学习阶段
 - 从已知类标（class label）的训练数据集中学习，生成分类器 $f(\textcolor{red}{x})$
 - 分类器又称为分类模型，可表示成分类规则、决策树或者数学公式
- 分类器应用（测试阶段）：
 - 用分类器 $f(\textcolor{red}{x})$ 来判断未知类标数据对象的类别

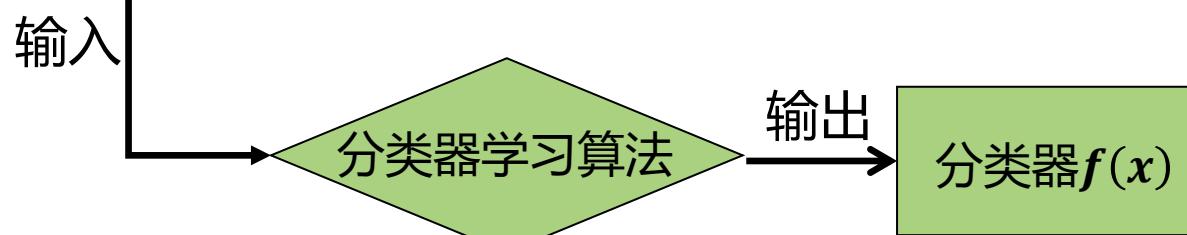
分类器的训练与应用



“贷款审批”的分类应用案例

序号	年龄	消费收入比	未偿还贷款	审批
1	中年	高	高	不通过
2	中年	一般	高	通过
3	中年	一般	较高	通过
4	中年	一般	低	通过
5	中年	一般	高	通过
6	老年	正常	低	通过
7	中年	超出	中	通过
8	中年	一般	较高	通过
9	青年	超出	低	通过
10	中年	正常	低	通过
11	中年	一般	较高	通过
12	中年	正常	低	不通过
13	中年	超出	中	不通过
14	中年	正常	高	不通过
15	中年	正常	低	不通过

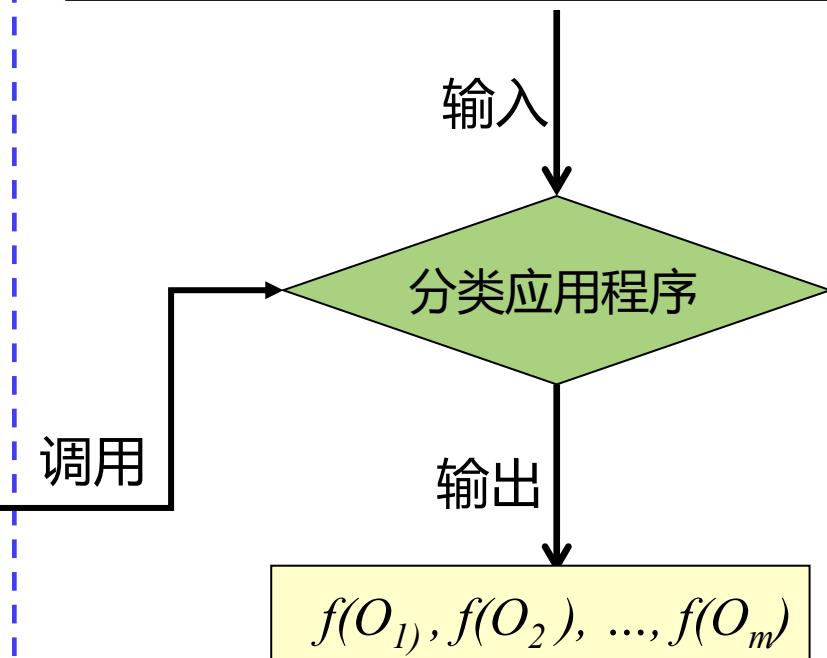
训练集D



训练阶段

测试集

序号	年龄	消费收入比	未偿还贷款	审批
1	中年	超出	高	?
2	青年	一般	中	?
3	老年	一般	较高	?
4	中年	正常	低	?



测试阶段

目录

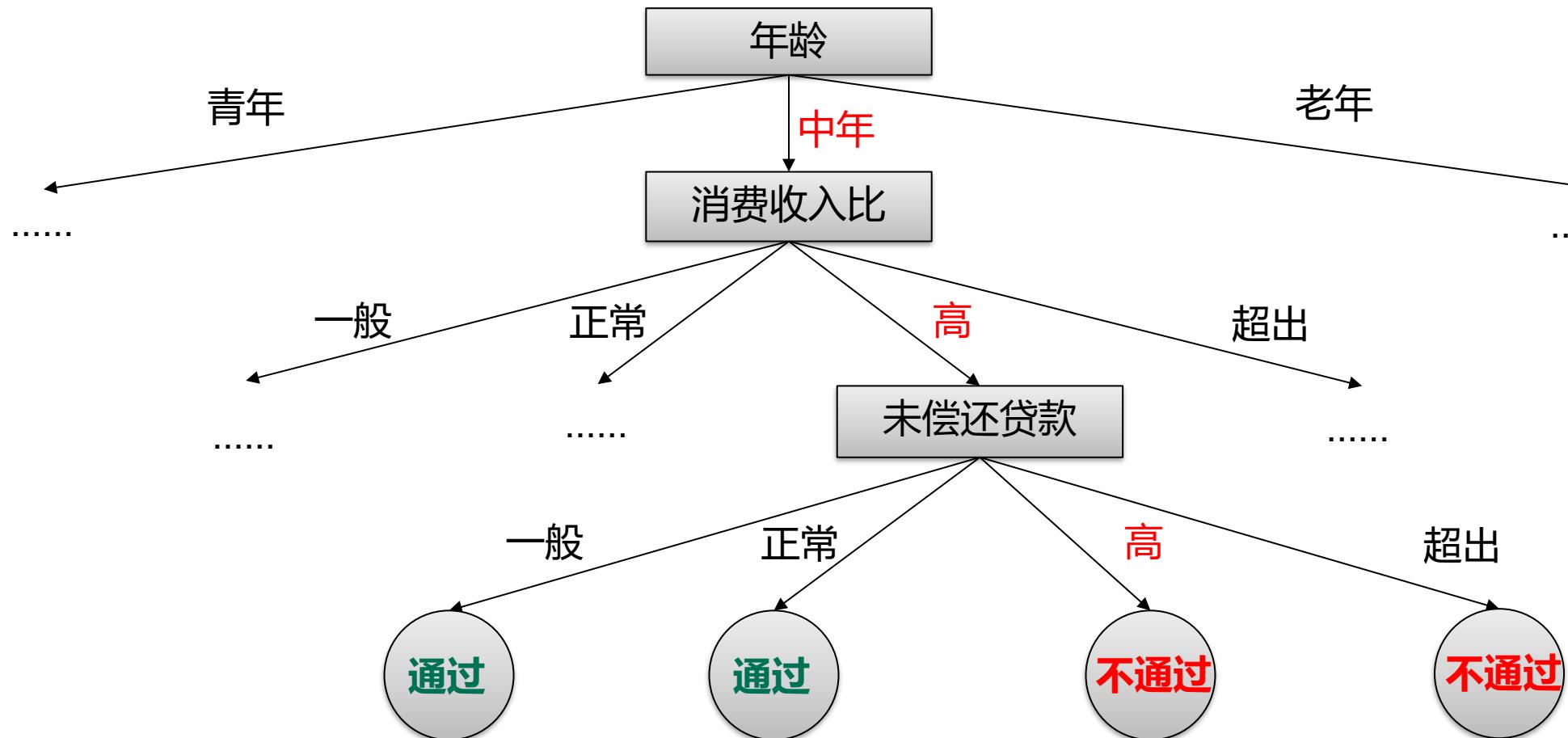
- 决策树分类的基本思想
- 分类决策树归纳算法
 - ID3
 - C4.5算法
 - CART算法
 - 决策树剪枝

决策树分类的基本思想

贷款审批的经验流程

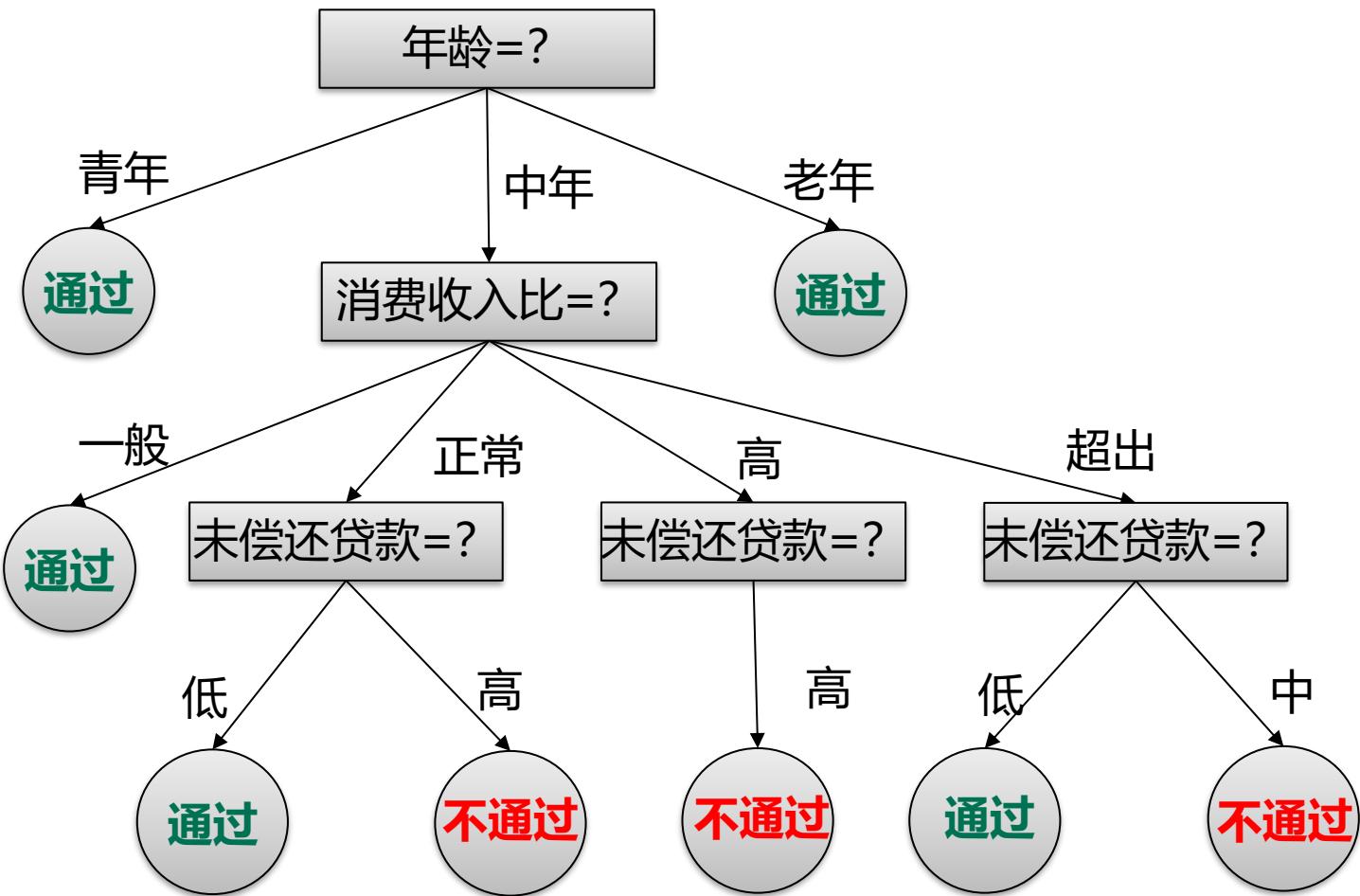
- 信贷审批员如何判断是否给一名借贷人房贷?

年龄	消费收入比	未偿还贷款
中年	高	高



决策树分类的基本思想：决策树归纳学习

- 决策树构建：对训练数据进行递归划分的过程



序号	年龄	消费收入比	未偿还贷款	审批
1	中年	高	高	不通过
2	中年	一般	高	通过
3	中年	一般	较高	通过
4	中年	一般	低	通过
5	中年	一般	高	通过
6	老年	正常	低	通过
7	中年	超出	中	通过
8	中年	一般	较高	通过
9	青年	超出	低	通过
10	中年	正常	低	通过
11	中年	一般	较高	通过
12	中年	正常	低	不通过
13	中年	超出	中	不通过
14	中年	正常	高	不通过
15	中年	正常	低	不通过

决策树分类的基本思想：决策树归纳学习

- 基础算法是贪心算法，算法要点：
 - 树以自顶向下的递归/分治的方式进行构建
 - 初始状态下，所有训练样本都处于树根的位置
 - 属性是用来对当前节点的训练数据集进行划分的（假定为离散属性，否则先离散化），即根据不同属性值划分
 - 选择对于当前节点“最优”的属性进行划分，划分过程递归进行
- 停止划分的条件
 - 给定节点的所有样本属于同一个类别
 - 没有剩余的属性可用于进一步分区 - 使用多数投票方法来对叶子进行分类
 - 没有多余的样本

ID3、C4.5、CART

信息增益

- 训练集D的信息不确定性程度 (**信息熵**) :

$$H(D) = - \sum_{i=1}^N p_i \log p_i = - \sum_{i=1}^N \frac{|D_i|}{|D|} \log \frac{|D_i|}{|D|}$$

- 根据属性A将D划分为 v 个子集后的信息熵 (**条件熵**) :

$$H(D|A) = \sum_{j=1}^v \frac{|D^{(j)}|}{|D|} H(D^{(j)})$$

- 信息增益**表示随机事件A发生后，对原数据D的不确定性减少程度

$$I(D; A) = H(D) - H(D|A)$$

为什么使用信息增益？

- 对于给定的训练集D与特征集A：
 - $H(D)$ 表示数据集D进行分类的不确定性
 - $H(D|A)$ 表示在特征A给定的条件下对数据集D进行分类的不确定性
 - $I(D;A) = H(D) - H(D|A)$ 表示由于特征A的引入使得对数据集D分类的不确定性减少的程度
- 由此，信息增益越大的特征，对训练集分类的不确定性减少越多，具有越强的分类能力

计算H(D)

$$H(D) = -\frac{10}{15} \log \frac{10}{15} - \frac{5}{15} \log \frac{5}{15} = 0.918$$

序号	年龄	消费收入比	未偿还贷款	审批
1	中年	高	高	不通过
2	中年	一般	高	通过
3	中年	一般	较高	通过
4	中年	一般	低	通过
5	中年	一般	高	通过
6	老年	正常	低	通过
7	中年	超出	中	通过
8	中年	一般	较高	通过
9	青年	超出	低	通过
10	中年	正常	低	通过
11	中年	一般	较高	通过
12	中年	正常	低	不通过
13	中年	超出	中	不通过
14	中年	正常	高	不通过
15	中年	正常	低	不通过

计算H(T|A=年龄)

$$H(D|A_1) = \frac{1}{15}H(T_{\text{青年}}) + \frac{13}{15}H(T_{\text{中年}}) + \frac{1}{15}H(T_{\text{老年}})$$

$$= \frac{1}{15} \times \left(-\frac{1}{1} \log \frac{1}{1} \right) +$$

$$\frac{13}{15} \times \left(-\frac{8}{13} \log \frac{8}{13} \right) - \left[\frac{5}{13} \log \frac{5}{13} \right] +$$

$$\frac{1}{15} \times \left(-\frac{1}{1} \log \frac{1}{1} \right)$$

$$= 0.833$$

序号	年龄	消费收入比	未偿还贷款	审批
1	中年	高	高	不通过
2	中年	一般	高	通过
3	中年	一般	较高	通过
4	中年	一般	低	通过
5	中年	一般	高	通过
6	老年	正常	低	通过
7	中年	超出	中	通过
8	中年	一般	较高	通过
9	青年	超出	低	通过
10	中年	正常	低	通过
11	中年	一般	较高	通过
12	中年	正常	低	不通过
13	中年	超出	中	不通过
14	中年	正常	高	不通过
15	中年	正常	低	不通过

计算H(D|A=消费收入比)、 H(D|A=未偿还贷款)

$$\begin{aligned}
 H(T|A_2) &= \frac{6}{15}H(T\text{一般}) + \frac{5}{15}H(T\text{正常}) + \frac{1}{15}H(T\text{高}) + \frac{3}{15}H(T\text{超出}) \\
 &= \frac{6}{15} \times \left(-\frac{6}{6} \log \frac{6}{6} \right) + \frac{5}{15} \times \left(-\frac{2}{5} \log \frac{2}{5} - \frac{3}{5} \log \frac{3}{5} \right) + \frac{1}{15} \times \left(-\frac{1}{1} \log \frac{1}{1} \right) + \\
 &\quad \frac{3}{15} \times \left(-\frac{2}{3} \log \frac{2}{3} - \frac{1}{3} \log \frac{1}{3} \right) = 0.507
 \end{aligned}$$

$$\begin{aligned}
 H(T|A_3) &= \frac{6}{15}H(T\text{低}) + \frac{2}{15}H(T\text{中}) + \frac{3}{15}H(T\text{较高}) + \frac{4}{15}H(T\text{高}) \\
 &= \frac{6}{15} \times \left(-\frac{4}{6} \log \frac{4}{6} - \frac{2}{6} \log \frac{2}{6} \right) + \frac{2}{15} \times \left(-\frac{1}{2} \log \frac{1}{2} - \frac{1}{2} \log \frac{1}{2} \right) + \\
 &\quad \frac{3}{15} \times \left(-\frac{3}{3} \log \frac{3}{3} \right) + \frac{4}{15} \times \left(-\frac{2}{4} \log \frac{2}{4} - \frac{2}{4} \log \frac{2}{4} \right) = 0.767
 \end{aligned}$$

序号	年龄	消费收入比	未偿还贷款	审批
1	中年	高	高	不通过
2	中年	一般	高	通过
3	中年	一般	较高	通过
4	中年	一般	低	通过
5	中年	一般	高	通过
6	老年	正常	低	通过
7	中年	超出	中	通过
8	中年	一般	较高	通过
9	青年	超出	低	通过
10	中年	正常	低	通过
11	中年	一般	较高	通过
12	中年	正常	低	不通过
13	中年	超出	中	不通过
14	中年	正常	高	不通过
15	中年	正常	低	不通过

计算I(D;A)

$$I(D; A_1) = H(D) - H(D|A_1) = 0.918 - 0.833 = 0.085$$

$$I(D; A_2) = H(D) - H(D|A_2) = 0.918 - 0.507 = 0.411$$

$$I(D; A_3) = H(D) - H(D|A_3) = 0.918 - 0.767 = 0.151$$

$$I(D; A_{max}) = I(D; A_2) = 0.411$$

最后，选择“消费收入比”的属性进行分支划分，划分分支为“一般、正常、高、超出”

序号	年龄	消费收入比	未偿还贷款	审批
1	中年	高	高	不通过
2	中年	一般	高	通过
3	中年	一般	较高	通过
4	中年	一般	低	通过
5	中年	一般	高	通过
6	老年	正常	低	通过
7	中年	超出	中	通过
8	中年	一般	较高	通过
9	青年	超出	低	通过
10	中年	正常	低	通过
11	中年	一般	较高	通过
12	中年	正常	低	不通过
13	中年	超出	中	不通过
14	中年	正常	高	不通过
15	中年	正常	低	不通过

连续属性的信息增益计算

- 对每个连续属性A:
 - 将A的值按递增顺序进行排序
 - 每个相邻值的中点被看做是可能的分裂点: $(\frac{a_i + a_{i+1}}{2})$
 - A的具有最小期望信息需求的点选做为A的分裂点
- 所有连续属性按以上方式计算出最佳分裂点，优中选优
- 如果选定A属性，进行**二叉划分**:
 - 对于分裂点 a_{split} ， D1是集合D中满足 $A \leq a_{\text{split}}$ 的元组合集合，而D2是集合D中满足 $A > a_{\text{split}}$ 的元组合集合

C4.5: 信息增益比

- ID3存在的bias问题：倾向于选择具有大量不同取值的属性
- 解决方法：信息增益比
 - 将信息增益进行归一化, 以克服选择过多属性值的bias

$$I_R(D; A) = \frac{I(D; A)}{SplitInfo_A(D)}$$

$$SplitInfo_A(D) = -\sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

基尼指数(CART, IBM IntelligentMiner)

- 数据集D包含n个类的样本, 基尼指数gini(D)定义为:
$$gini(D) = 1 - \sum_{j=1}^n p_j^2$$
- 二叉划分策略: 如果属性A的二元划分将数据集D划分成 D_1 and D_2 , 则给定该划分, D的基尼指数gini(D)定义为:
$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$
- “不纯度”降低为:
$$\Delta gini(A) = gini(D) - gini_A(D)$$
- 选择产生最小基尼指数 $gini_{split}(D)$ (或者最大化不纯度降低) 的属性作为分裂属性

CART: 离散属性的二叉划分问题

信息增益、信息增益率和基尼系数对比

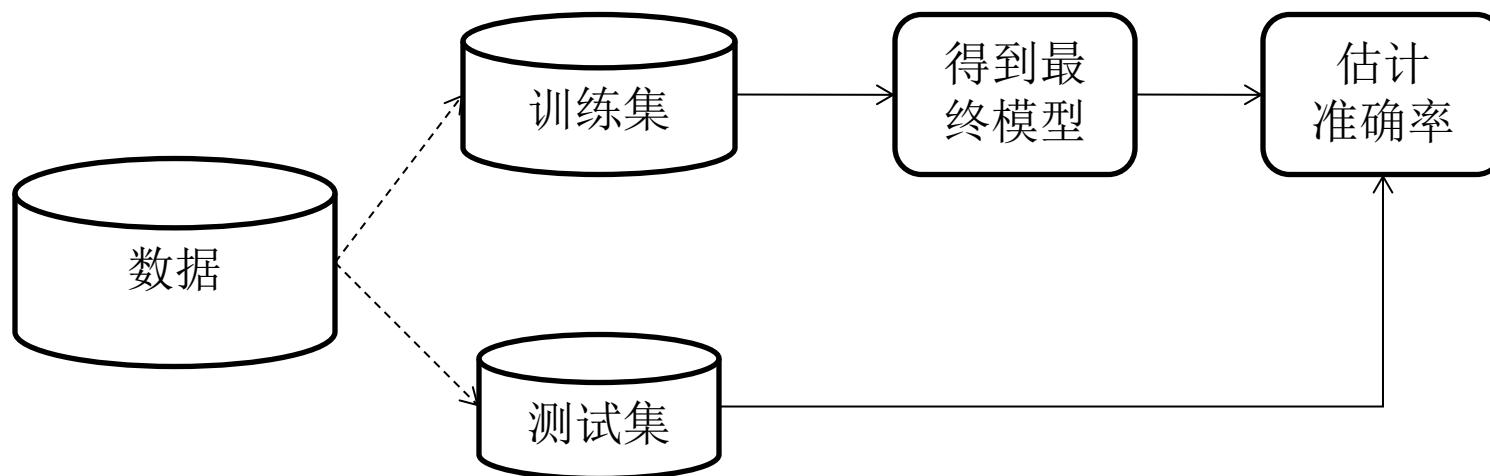
- 总得来说，这三种度量都能得到良好的结果。但是：
 - 信息增益：
 - ✓ 倾向于多值属性
 - 增益率：
 - ✓ 倾向于产生不平衡的划分，其中一个分区比其他分区小得多
 - 基尼指数：
 - ✓ 倾向于多值属性
 - ✓ 当类的数量很大时会有困难
 - ✓ 倾向于导致相等大小的分区和纯度

- 思考：能否用卡方 χ^2 作为属性选择方法？

决策树剪枝

分类器性能的评估

- 乐观估计：基于训练集误差
- 独立测试样本评估：将给定的数据随机划分为两个独立的集合
 - **训练集**（例如 $2/3$ ）用于模型构建
 - **测试集**（例如 $1/3$ ）用于准确率估计



分类器准确性度量指标

- 分类器M的准确率 $\text{Acc}(M)$: 分类器M正确分类的样本所占比例
➤ 分类器M的错误率(误分类率) = $1 - \text{acc}(M)$

欠拟合与过拟合的基本概念

- 欠拟合
- 过拟合

决策树的过拟合问题和剪枝方法

- 过拟合： 决策树归纳可能过度拟合训练数据
 - 由于数据中的噪声和离群点，许多分枝反映的是训练数据中的异常
 - 对未知样本分类精度低
 - 通常表现为“复杂”的树
- 两种常用的方法可以避免过拟合：
 - 先剪枝： 提前停止树的构建
 - 后剪枝： 从“完全生长”的树剪去子树——产生一个渐进的剪枝树的集合、

预剪枝策略

- 在决策树建树过程中便对决策树的生长进行控制，一旦符合下列条件，决策树便停止生长
 - 限制决策树的最高高度
 - 设定叶子节点正确划分率
 - 设定叶子节点最少样本数量
 -

后剪枝策略

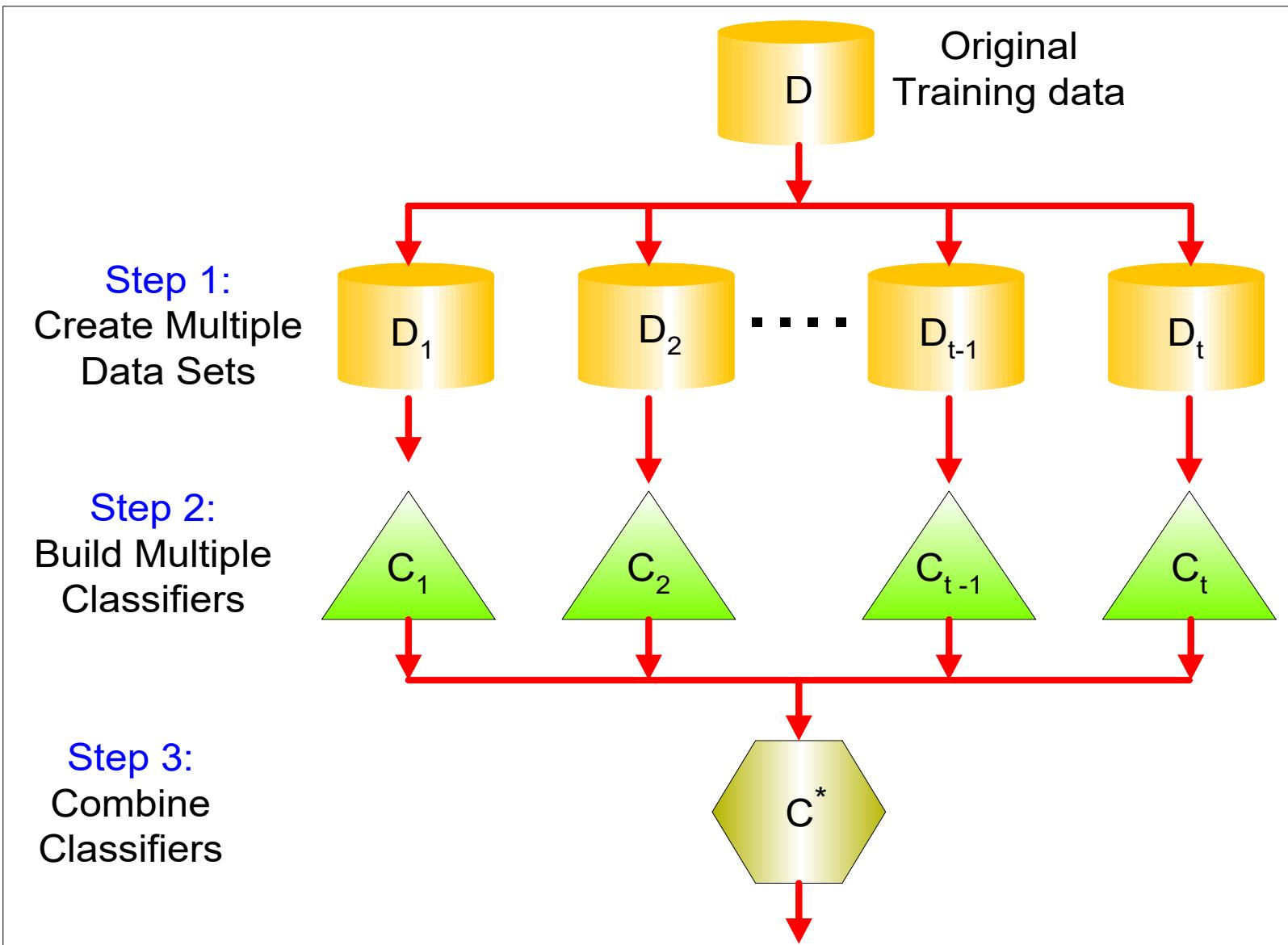
- 决策树生成完成后，根据一定的条件判断某些子树的过拟合程度，动态进行修剪，从而限制决策树的最高高度
 - 如：误差降低剪枝
- 使用独立于训练集（用于建立未剪枝树）的样本集

集成学习分类方法

基于集成学习的分类方法

- 集成学习 (ensemble learning)
- 从训练数据构建分类器集合
- 综合多个分类器的预测结果来预测新实例的分类预测结果

总体思路



集成学习的关键问题

- 强度 (strength)
- 相关性 (correlation)
- Tradeoff between strength and correlation

集成方法的例子

- 如何生成一个集成分类器?
 - 装袋 (Bagging)
 - 提升 (Boosting)
 - 梯度提升 (Gradient Boosting)
 - 堆叠 (Stacking)

装袋 (Bagging)

- 有放回抽样

Original Data	1	2	3	4	5	6	7	8	9	10
Bagging (Round 1)	7	8	10	8	2	5	10	10	5	9
Bagging (Round 2)	1	4	9	1	2	3	2	7	3	2
Bagging (Round 3)	1	8	5	10	5	5	9	6	3	7

- 为每个自助样本建立分类器
- 每个样本被抽取的概率是相等的

随机森林

- 专门为决策树分类器设计的集成方法
- 随机森林长出许多分类树 (名字由来)
- 未修剪的决策树集合
- 每个基础分类器分类一个 “新的” 向量
- 森林选择表决投票最多的分类结果

随机森林

- 随机性的两个来源介绍：“装袋” 和 “随机输入向量”
 - 每棵树都是使用训练数据的bootstrapping样本来生长的
 - 在每个节点上，最佳的分裂是从 m_{try} 个变量中选择，而不是所有变量

随机森林

