第3章 决策树

本章目标

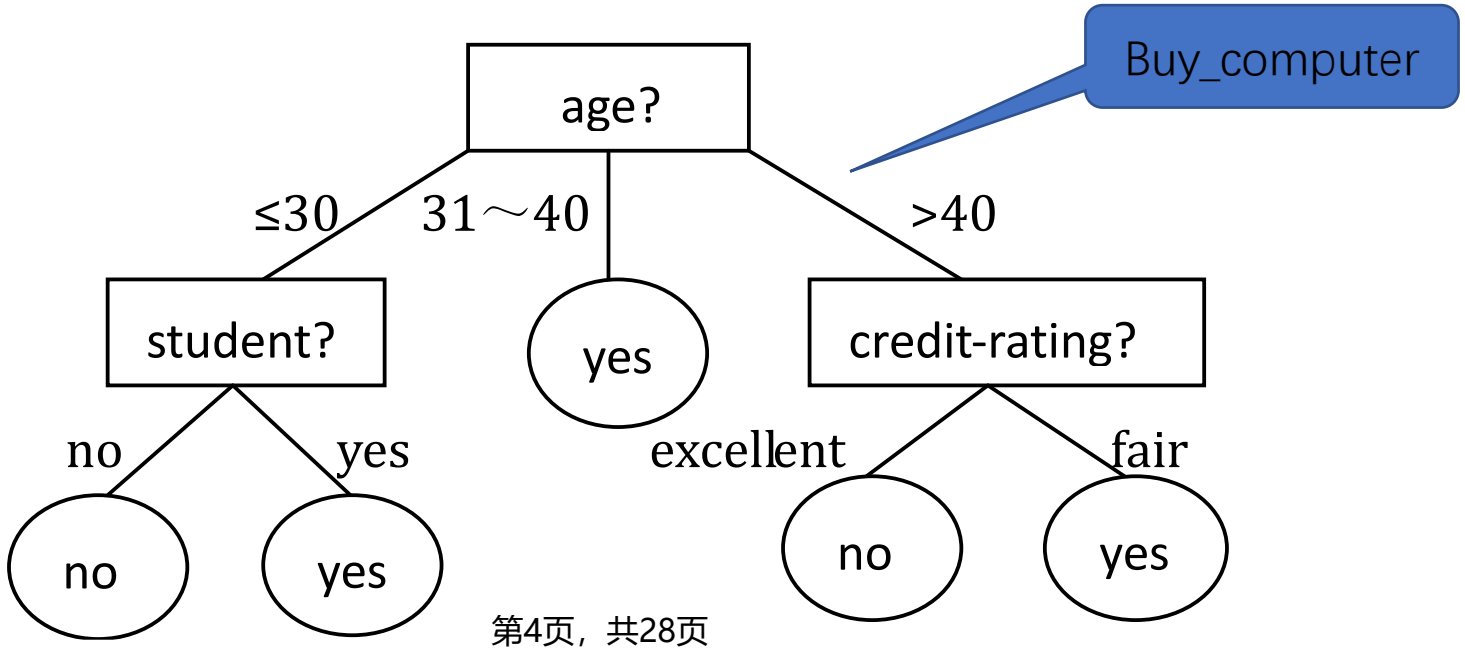
- 了解决策树的概念;
- 了解C4.5决策树建立过程、关键技术、和决策树规则;
- 了解其他决策树算法。

决策树学习

- 决策树（Decision Tree）是数据挖掘中最常用的一种分类和预测技术。
- 它从一组无次序、无规则的事例中推理出决策树表示形式的分类规则。
- 构建好的决策树呈树形结构，可以认为是if-then规则的集合，主要优点是模型具有可读性，分类速度快。

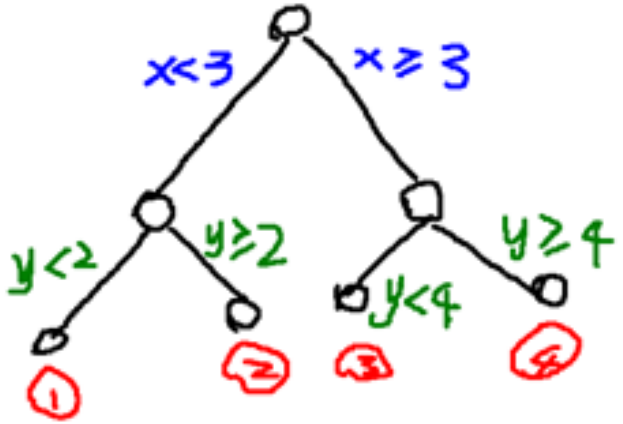
决策树学习

- 决策树模型是一个树状结构，树中每个节点表示分析对象的某个属性，每个分支表示这个属性的某个可能的取值。
- 决策树分类方法采用自顶向下的递归方式，在决策树的内部节点进行属性值的比较，根据不同的属性值判断从该节点向下的分支，在决策树的叶节点得到结论。

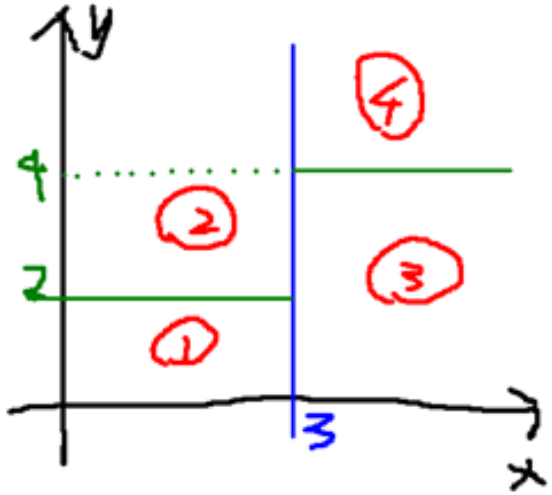


决策树学习

- 决策树实际上是将空间用超平面进行划分的一种方法，每次分割的时候，都将当前的空间一分为二，比如说下面的决策树：



就是将空间划分成下面的样子：



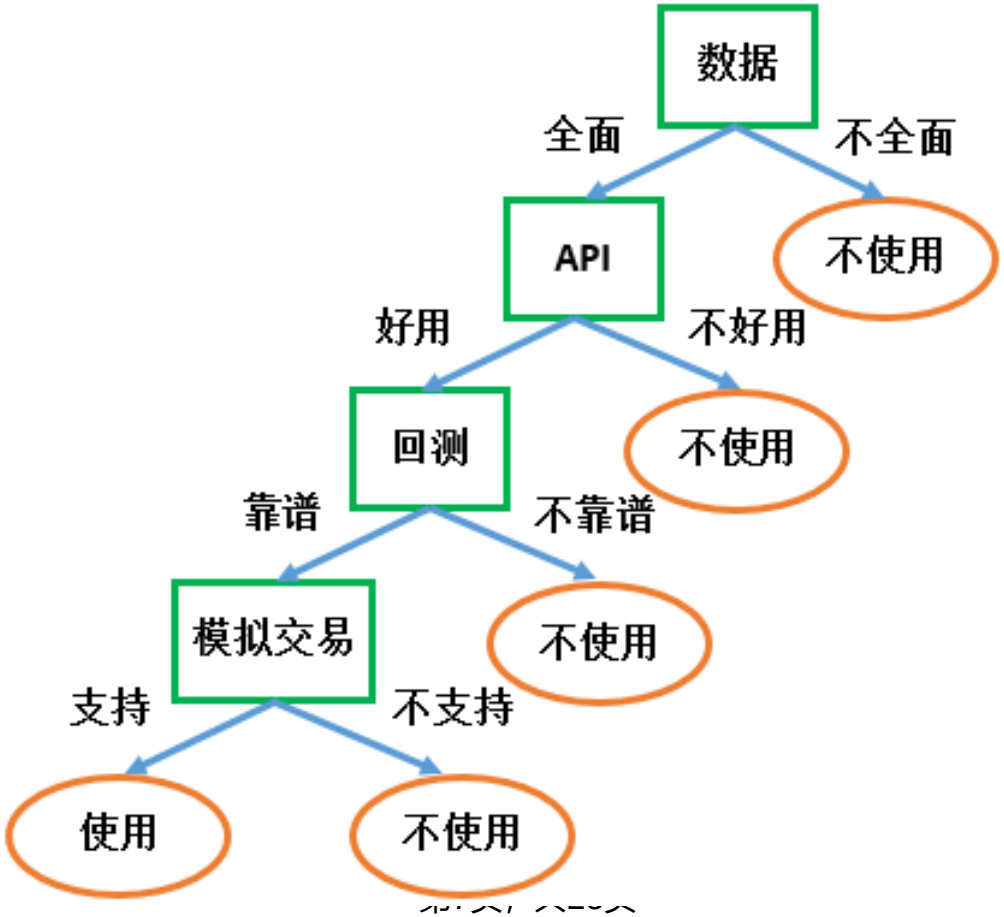
这样使得**每一个叶子节点都是在空间中的一个不相交的区域**，在进行决策的时候，会根据输入样本每一维特征的值，一步一步往下，最后使得样本落入N个区域中的一个（假设有N个叶子节点）

决策树学习

- 用选择量化工具的过程形象的展示一下决策树的构建。假设现在要选择一个优秀的量化工具来帮助我们发掘好的股票，怎么选呢？
  - 第一步：看看工具提供的数据是不是全面。
  - 第二步：看看工具提供的API是不是好用。
  - 第三步：看看工具的回测过程是不是靠谱。
  - 第四步：看看工具支不支持模拟交易，光回测只是能让你判断策略在历史上有用没有，正式运行前起码需要一个模拟盘吧。
- 这样，通过将“数据是否全面”，“API是否易用”，“回测是否靠谱”，“是否支持模拟交易”将市场上的量化工具贴上两个标签，“使用”和“不使用”。

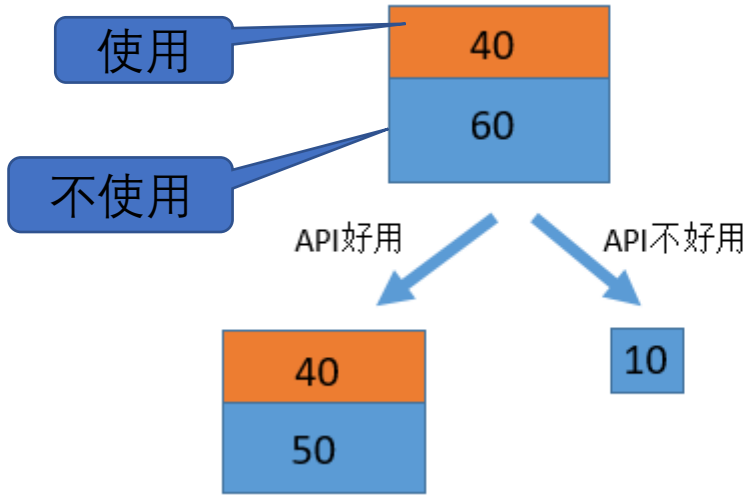
决策树学习

- 上面就是一个决策树的构建，逻辑可以用下图表示：



决策树学习

- 可以看到，决策树的主要工作，就是选取特征对数据集进行划分，最后把数据贴上两类不同的标签。如何选取最好的特征呢？



决策树学习

- 在现实应用中，数据集往往不能达到上述“是否支持模拟交易”的分类效果。所以我们用不同的准则衡量特征的贡献程度。主流准则的列举3个：
  - CART算法（Breiman等人于1984年提出）利用基尼指数最小化准则进行特征选择。
  - ID3算法（J. Ross Quinlan于1986年提出），采用信息增益最大的特征；
  - C4.5算法（J. Ross Quinlan于1993年提出）采用信息增益比选择特征；

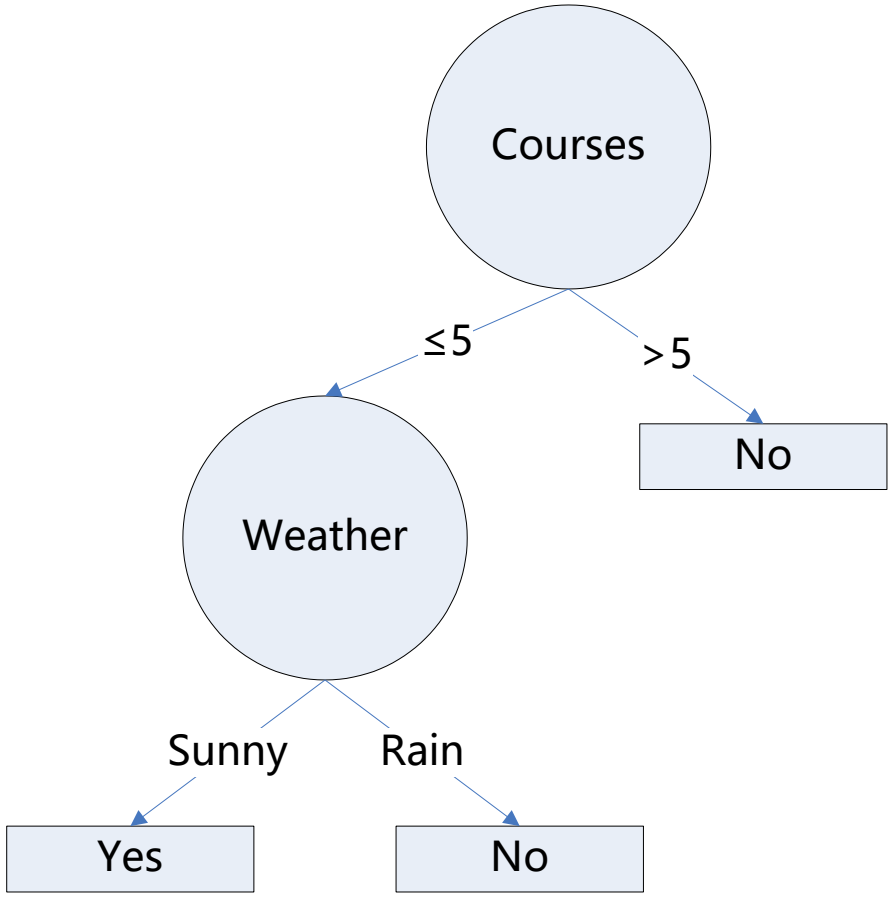
【例2.1】

给定如表2.1所示的数据集T，建立一棵决策树，用于预测某个学生是否决定去打篮球。

表2.1 一个假想的打篮球数据集

序号	Weather	Temperature/°C	Courses	Partner	Play
1	Sunny	20~30	4	Yes	Yes
2	Sunny	20~30	4	No	Yes
3	Rain	10~0	1	Yes	Yes
4	Sunny	30~40	5	Yes	Yes
5	Rain	20~30	8	No	No
6	Sunny	-10~0	5	Yes	Yes
7	Sunny	-10~0	7	No	No
8	Rain	20~30	2	Yes	Yes
9	Rain	20~30	6	Yes	No
10	Sunny	10~20	6	Yes	No
11	Rain	10~20	3	No	No
12	Rain	10~20	1	Yes	No
13	Sunny	10~20	8	Yes	No
14	Sunny	0~10	3	Yes	Yes
15	Rain	0~10	2	Yes	No

决策树



- 使用15个实例进行有训练，其中Weather、Temperature、Courses和Partner作为输入属性，Play作为输出属性。

图2.1 打篮球决策树