

01 绪论

机器学习：通过计算，利用经验来改善系统自身性能

基本术语

- 数据
 1. 数据集：100个西瓜
 2. 样本：1个西瓜
 3. 特征向量：
 1. 样本空间
 2. 大小、颜色、振幅
 3. 维度
 4. 属性：颜色（一个向量）
 - 利用某种学习算法
 - 得到模型
 1. 分类（有监督学习）：
 1. 二分类：Y的正负；瓜摘还是不摘
 2. 多分类：买哪种西瓜
 2. 回归（有监督学习）：
 1. $Y=R$ (实数集)：已知各段时间西瓜的价格，问什么时间买瓜最好
 3. 聚类（无监督学习）：
 1. 我们不知道要怎么分，让机器分类
 2. 每个组称为“簇”(cluster)
 - 进行预测
 1. 测试
 2. 测试样本
 3. 泛化能力：预测到没见过的数据(一般泛化准确度达到90%以上才行)
-

假设空间

科学推理手段：

1. 归纳：特殊到一般
 1. 狭义：从数据中得到概念
 1. bool概念：好瓜 = (色泽=?) ^ (根蒂=?) ^ (敲声=?) 各种情况的分类组合得到假设
 2. 广义：从样本中学习
 2. 演绎：一般到特殊
-

归纳偏好

由一个数据集训练出了不同模型，如何选择模型

奥卡姆剃刀原则：选最简单的那个