1.1 引言

略

1.2 基本术语

1.2.1 数据集相关的基本概念

> 假定我们收集了一批关于西瓜的数据，例如(色泽=青绿;根蒂=蜷缩;敲声=浊响)，(色泽=乌黑;根蒂=稍蜷;敲声=沉闷)，(色泽=浅白;根蒂=硬挺;敲声=清脆)，……，每对括号内是一条记录，"="，意思是"取值为" .

1. **数据集(data set)** 这组记录的集合称为一个"数据集" (data set).
2. **示例(instance)或样本(sample)** 其中每条记录是关于一个事件或对象(这里是一个西瓜)的描述 ,称为一个"示例" (instance) 或"样本" (sample).一个示例也可以称为一个**特征向量(feature vector)**.
3. **属性(attribute)或特征(feature)** 反映事件或对象在某方面的表现或性质的事项，例如"色泽" "根蒂" "敲声"
4. **属性值(attribute value)** 属性上的取值，例如"青绿" "乌黑"，称为")副主值" (attribute value).
5. **属性空间(attribute space)、 样本空间(sample space)或输入空间** 属性张成的空间称为"属性空间" (attribute space) 、 "样本空间" (sample space)或"输入空间" .

> 一般的,令$D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$表示包含 m 个示例的数据集, 每个示例由 d 个属性描述(例如上面的西瓜数据使用了 3 个属性)，则每个示例 $\boldsymbol{x}_i = (x_{i1}; x_{i2}; \ldots; x_{id})$ 是 d 维样本空间 $\mathcal{X}$ 中的一个向量， $\boldsymbol{x}_i \in \mathcal{X}$， 其中 x_{ij} 是 $\boldsymbol{x}_i$ 在第 j 个属性上的取值, d 称为样本 $\boldsymbol{x}_i$ 的"维数" (dimensionality).

1.2.2 训练过程中的相关概念

> 从数据中学得模型的过程称为"学习" (learning)或"训练" (training), 这个过程通过执行某个学习算法来完成.

1. **训练数据(training data)** 训练过程中使用的数据称为"训练数据" (training data)
2. **训练样本(training sample)** 其中每个样本称为一个"训练样本" (training sample)
3. **训练集(training set)** 训练样本组成的集合称为"训练集" (training set)
4. **假设(hypothesis)** ***学得模型***对应了关于数据的某种潜在的规律，因此亦称**"假设" (hypothesis)**.学习过程就是为了找出或逼近真相. 有时将模型称为**"学习器" (learner)**

1.2.3 label相关概念

> 这里关于示例结果的信息，例如"好瓜"，称为**"标记" (label)**; 拥有了标记信息的示例，则称为**"样例" (example)**. 一般地，用 $(\boldsymbol{x}_i, y_i)$ 表示第 i 个样例, 其中 $y_i \in \mathcal{Y}$ 是示例 x_i 的标记, $\mathcal{Y}$ 是所有标记的集合, 亦称"标记空间"(label space)或"输出空间"

1. **标记(label)** 关于示例结果的信息，例如"好瓜"，称为"标记" (label)
2. **样例(example)** 拥有了标记信息的示例，则称为"样例" (example).

预测类型	学习任务名称	分类
离散值	分类(classification)	"二分类"和"多分类"
连续值	回归(regression)	

1.2.4 测试相关概念

1. **测试(testing)** 学得模型后，使用其进行预测的过程称为"测试" (testing) .
2. **测试样本(testing sample)** 被预测的样本称为"测试样本" (testing sample). 例如在学得 f 后，对测试例 x_i , 可得到其预测标记 $y = f(x)$

1.2.5 学习任务的划分

根据训练数据是否拥有标记信息，学习任务可大致划分为两大类**"监督学习" (supervised learning)** 和**"无监督学习" (unsupervised learning)** ，分类和回归是前者的代表，而聚类则是后者的代表.

1.2.6 泛化

1.**泛化(generalization)** 学得模型适用于新样本的能力，称为"泛化" (generalization)能力

> 具有强泛化能力的模型能很好地适用于整个样本空间,一般而言，训练样本越多，我们得到的关于 D 的信息越多，这样就越有可能通过学习获得具有强泛化能力的模型.

1.3 假设空间

1.3.1 归纳和演绎

归纳 (induction)与横绎 (deduction是科学推理的两大基本手段.前者是**从特殊到一般**的"泛化" (generalization)过程，即从具体的事实归结出一般性规律;后者则是**从一般到特殊**的"特化" (specialization)过程，即从基础原理推演出具体状况.

> "从样例中学习"是一个归纳的过程, 因此亦称**"归纳学习" (inductive learning)**

1.3.2 归纳学习

1. 广义的归纳学习大体相当于从样例中学习
2. 狭义的归纳学习则要求从训练数据中学得概念 (concept) ， 因此亦称为"概念学习"或"概念形成" .概念学习中最基本的是布尔概念学习, 即对"是" "不是"这样的可表示为 0/1 布尔值的目标概念的学习.

1.3.3 假设空间

1. 学习过程看作一个在所有假设(hypothesis)组成的空间中进行搜索的过程，搜索目标是找到与训练集"匹配"(fit)的假设，即能够将训练集中的瓜判断正确的假设.
2. 假设的表示一旦确定，假设空间及其规模大小就确定了.
3. 例: 设空间由形如"(色泽=?)∧(根蒂=?) ∧(敲声=?)"的可能取值所形成的假设组成, 加上通配符"*",和空集∅,共有 $4 \times 3 \times 3 + 1 = 37$

1.3.4 版本空间

可能有多个假设与训练集一致，即存在着一个与训练集一致的"假设集合" ,称之为"版本空间" (version space) .

表 1.1 西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

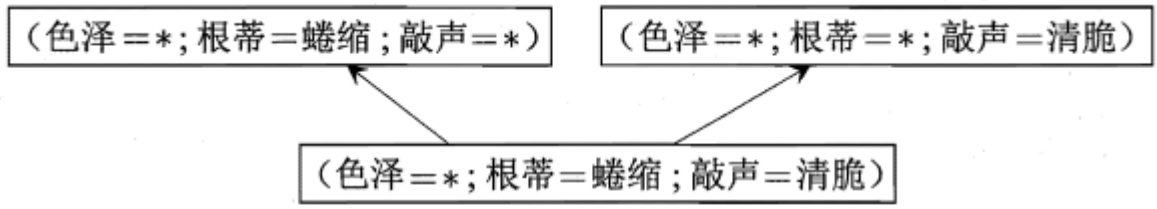

图 1.2 西瓜问题的版本空间

1.4 归纳偏好

1.4.1 归纳偏好的概念

机器学习算法在学习过程中对某种类型假设的偏好，称为"归纳偏好" (inductive bias),或简称为"偏好"

> 任何一个有效的机器学习算法必有其归纳偏好，否则它将被假设空间中看似在训练集上"等效"的假设所迷惑，而无法产生确定的学习结果.

1.4.2 奥卡姆剃刀

1. 概念: "奥卡姆剃刀" (Occam's razor)是一种常用的、自然科学研究中最基本的原则，即"若有多个假设与观察一致，则选最简单的那个" .
2. 奥卡姆剃刀也并非唯一可行的原则

1.4.3 没有免费的午餐(NFL)

1. 无论学习算法 $\mathfrak{L}_a$ 多聪明、学习算法 $\mathfrak{L}_b$ 多笨拙，它们的期望性能竟然相同!这就是"没有免费的午餐"定理 (No Free Lunch Theorem，简称 NFL.
2. NFL 定理有一个重要前提:所有"问题"出现的机会相同、或所有问题同等重要 .
3. 但实际上,我们只关注自己正在试图解决的问题，希望为它找到一个解决方案，至于这个解决方案在别的问题、甚至在相似的问题上是否为好方案，我们并不关心.

> 脱离具体问题，空泛地谈论"什么学习算法更好"毫无意义，因为若考虑所有潜在的问题,则所有学习算法都一样好.要谈论算法的相对优劣，必须要针对具体的学习问题;在某些问题上表现好的学习算法，在另一些问题上却可能不尽如人意，学习算法自身的归纳偏好与问题是否相配，往往会起到决定性的作用.

1.5 发展历程

略

1.6 应用现状

略