决策树

1、机器学习方法的一种，解决分类问题
2、信息熵：度量样本集合纯度的指标，其值越大，纯度提升越大
3、增益率，c4.5决策树算法，先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率高的。
4、基尼系数，CART决策树：反映从数据集中随机抽取两个样本，其类别标记不一样的概率，其值越小，纯度越高
5、剪枝处理（预剪枝、后剪枝）
（1）、预剪枝：在决策树生成过程，对每个结点在划分前进行估计，若划分不能带来决策树泛化性能提升，则停止划分并将当前结点标记为叶结点。容易欠拟合。
（2）、后剪枝：先生成一棵完整的决策树，然后自底向上对非叶子结点进行考察，若剪枝能带来决策树泛化性能提升，则将该子树替换为叶结点。
6、连续值处理：连续属性离散化
7、缺失值处理：若属性取值已知，则划入与其取值对应的子结点；若属性未知，则加权划入所有子结点。
8、多变量决策树：不是为每个非叶结点寻找一个最优划分属性，而是试图建立一个合适的线性分类器。

基本术语

1、样本、示例，属性、特征，属性值、，属性空间、样本空间、输入空间
2、维数：样本的属性个数
3、训练集、训练数据，测试集、测试样本，标记
4、标记空间、输出空间：标记的合集
5、分类：预测离散值，如好瓜、坏瓜；二分类：正类、反类；多分类
6、回归：预测连续值，如西瓜成熟度0.95、0.37
7、聚类：将无标记信息的训练集分成若干组，每组称为一个‘簇‘，自动形成的簇对应一些潜在概念的划分，这样的概念事先不知道。
8、监督学习：分类、回归：无监督学习：聚类
9、泛化能力：模型适用于新样本的能力。
10、假设空间、版本空间：假设集合
11、归纳偏好：对某种类型假设的偏好。奥卡姆剃刀原则

模型评估与选择

1、错误率：分类错误的样本占样本总数的比例；精度

2、误差：学习器的实际预测输出与样本的真实输出之间的差异

3、训练误差、经验误差：学习器在训练集上的误差

4、泛化误差：在新样本上的误差

5、过拟合：学习器把训练样本自身的一些特点当作所有潜在样本都会具有的一般性质，不容易克服；欠拟合：对训练样本的一般性质尚未学好，容易克服

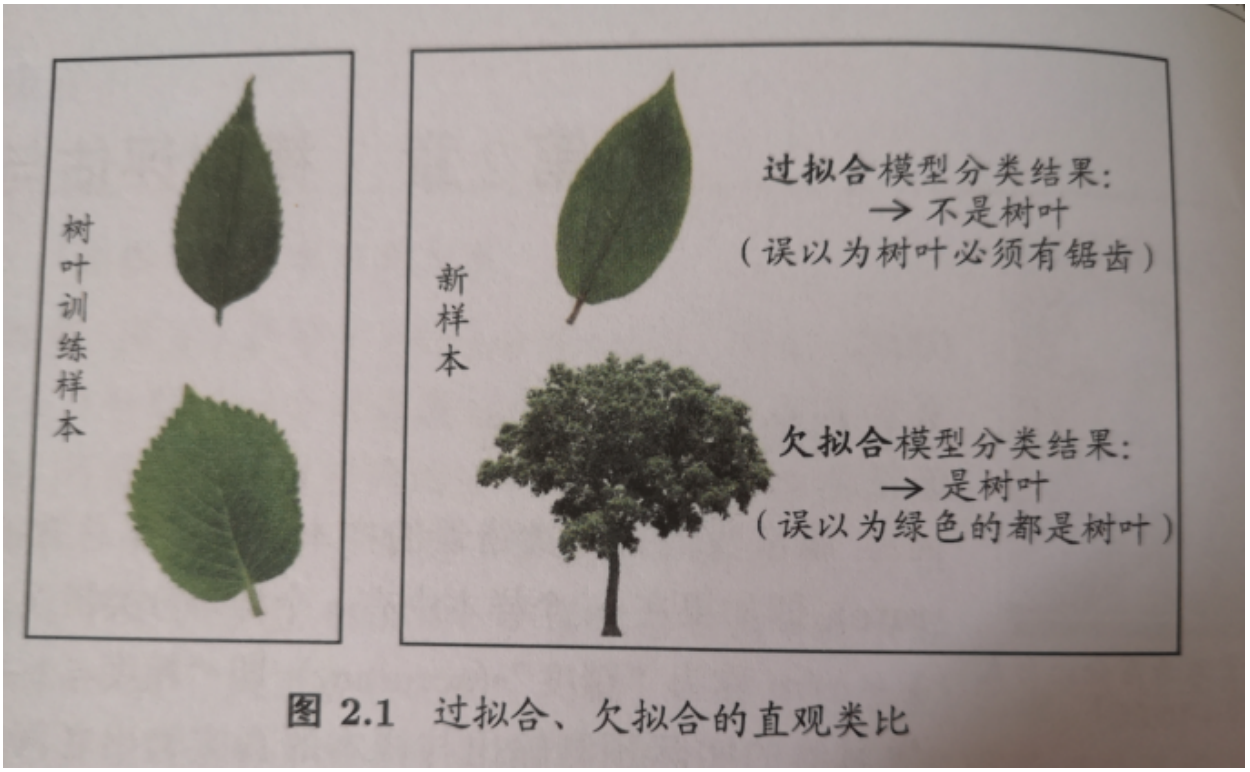

图 2.1 过拟合、欠拟合的直观类比

6、模型评估方法

（1）留出法：直接将数据集划分为两个互斥的集合，一个训练集，一个测试集，用测试集评估其测试误差，测试误差作为泛化误差的近似。一般采用若干次随机划分、重复进行试验评估后取平均值作为评估结果。

（2）交叉验证法：将数据集划分为k个互斥子集，k-1个子集的并集作为训练集，余下子集作为测试集。最终返回k个测试结果的均值，所以交叉验证法又称“k折交叉验证”。随机使用不同划分重复p次，最终评估结果是这p次k折交叉验证结果的均值。k=样本数时，称为留一法。

（3）自助法：假设数据集包含m个样本，有放回的采样，重复执行m次，得到包含m个样本的训练集

7、调参：用测试集上的判别效果来估计模型在实际应用时的泛化能力，用验证集上的性能来进行模型选择和调参。

线性模型

1、试图学得一个通过属性的线性组合来进行预测的函数。

2、对数线性回归，广义线性模型，对数几率回归

3、线性判别分析（LDA）

4、多分类学习，编码，解码

5、类别不平衡问题的处理：欠采样，过采样，阈值移动

6、对离散属性，若属性间存在序关系，通过连续化将其转化为连续值，例如三值属性高度的取值高、中、低可转化为{1.0,0.5,0.0}；若属性值间不存在序关系，假定有k个属性值，通常转化为k维向量，例如属性瓜类的取值西瓜、南瓜、黄瓜可转化为（0,0,1），（0，1,0），（1,0,0）。

神经网络

1、N-P神经元模型

2、感知机：两层神经元组成（输入层、计算层）

3、多层前馈神经网络：每层神经元与下一层神经元全互连，神经元之间不存在同层连接，也不存在跨层连接。

4、多层网络：包含隐层。

5、误差逆传播（标准BP算法）

6、积累BP算法

7、缓解BP网络的过拟合：早停，正则化

8、全局最小与局部最小

9、RBF网络：一种单隐层前馈神经网络

10、ART网络：竞争型学习是神经网络中一种常用的无监督学习策略，使用该策略时，网络的输出神经元相互竞争，每一时刻仅有一个竞争获胜的神经元被激活，其他神经元的状态被抑制。ART网络是竞争型学习的代表。ART可进行增量学习或在线学习。

11、SOM网络：一种竞争型的无监督神经网络，将高维空间中相似的样本点映射到网络输出层中的近邻神经元。

12、级联相关网络

13、Elman网络：递归神经网络允许网络中出现环形结构，让一些神经元的输出反馈回来作为输入信号。Elman网络是最常用的递归神经网络之一。

14、Boltzmann机：神经网络中有一类模型是为网络状态定义一个能量，能量最小化时网络达到理想状态，而网络的训练就是在最小化这个能量函数。Boltzmann机就是一种基于能量的模型。

15、RBN：受限Boltzmann机

16、深度学习：很深层的神经网络。特征学习、表示学习。

支持向量机

1、支持向量：距离超平面最近的几个特殊训练样本点

2、间隔：两个异类支持向量到超平面的距离之和

3、支持向量机（SVM）基本模型：最大化间隔

4、对偶函数

5、核函数

6、硬间隔：要求所有样本均满足约束：软间隔：允许某些样本不满足约束

7、替代损失函数：hinge损失，指数损失，对率损失

8、支持向量回归（SVR）:容忍预测值与真实值有一定误差。

9、核方法：基于核函数的学习方法，引入核函数将线性学习器拓展为非线性学习器。

10、解决分类问题

贝叶斯分类器

1、贝叶斯决策论：基于概率和误判损失来选择最优的类别标记。

2、后验概率，类先验概率，类条件概率（似然）

3、极大似然估计（MLE）:根据数据采样来估计概率分布参数的经典方法。

4、朴素贝叶斯分类器：采用了“属性条件独立性假设”，即假设每个属性独立地对分类结果发生影响。

5、半朴素贝叶斯分类器：采用独依赖估计策略。

6、贝叶斯网：亦称信念网，借助有向无环图刻画属性间的依赖关系，用条件概率表来描述属性的联合概率分布。

7、结构：同父结构，V型结构，顺序结构

8、道德图

9、评分搜索

10、贝叶斯网训练好之后能通过一些属性变量的观测值来推测其他属性变量的取值。

11、推断：通过已知变量观测值来推测待查询变量，已知变量观测值称为证据

12、EM算法：常用的估计参数隐变量的利器

集成学习

1、通过构建并结合多个学习器来完成学习任务，也称为多分类器系统，基于委员会的学习。

2、集成学习方法分两大类：

（1）个体学习器间存在强依赖关系，必须串行生成的序列化方法，代表是Boosting

（2）个体学习器间不存在强依赖关系，可同时生成的并行化方法，代表是Bagging和随机森林

3、弱学习器：泛化性能略优于随机猜测的学习器，例如二分类问题上精度略高于50%的分类器。

4、基学习器，基学习算法，组件学习器。

5、Boosting：将弱学习器提升为强学习器的算法。

6、AdaBoost算法

7、Bagging:直接基于自助采样法

8、随机森林（FR）:Bagging的一个拓展变体。以决策树为基学习器构建集成，在决策树的训练过程引入随机属性选择。

9、结合策略：平均法，投票法，学习法

10、多样性度量：用于度量集成中个体分类器的多样性。

11、多样性增强：数据样本扰动，输入属性扰动，输出表示扰动，算法参数扰动

聚类

1、性能度量亦称有效性指标。外部指标：将聚类结果与某个参考模型进行比较：内部指标：直接考察聚类结果不利用任何参考模型。

2、距离计算

3、原型聚类：k均值算法，学习向量量化（LVQ），高斯混合聚类，密度聚类，层次聚类

降维与度量学习

1、k近邻学习：基于某种距离度量找出训练集中与其最靠近的k个训练样本，然后基于这k个“邻居”的信息来进行预测。分类投票法，回归平均法。

2、维数灾难：高维情形下出现的数据样本稀疏，距离计算困难等问题。

3、低维嵌入

4、降维方法：主成分分析（PCA），核化线性降维，流行学习（等度量学习，局部线性嵌入）
5、度量学习：直接学习一个合适的距离度量。
6、懒惰学习，急切学习。

特征选择与稀疏学习
1、特征：相关特征，无关特征
2、三类特征选择方法：过滤式，包裹式，嵌入式。
3、过滤式方法先对数据集进行特征选择，然后再训练学习器，特征选择过程与后续学习器无关。
4、包裹式特征选择就是为给定学习器选择最有利于器性能量身定做的特征子集。
5、嵌入式特征选择在学习器训练过程中自动地进行了特征选择。
6、字典学习、稀疏编码：为普通稠密表达的样本找到合适的字典，将样本转化为合适的稀疏矩阵表示形式。
7、压缩感知：基于稀疏性从少量观测中恢复原信号。