VC维

H的VC维表示为VC(H) ，指能够被H分散的最大集合的大小。若H能分散任意大小的集合，那么VC(H)为无穷大。

一般地,在r 维空间中,线性决策面的VC维为r+1。

几个指标

范数

生成模型与判别模型

判别式模型：给定x，可通过直接建模P(c|x)来预测c。比如决策树，BP神经网络，SVM支持向量机，条件随机场CRF，区分度训练，逻辑斯蒂回归，最近邻KNN，线性判别分析，Boosting，线性回归，CART，高斯过程

生成式模型：先对联合概率P(x,c)建模，然后得到P(c|x)。比如贝叶斯分类器，混合高斯模型，隐马尔科夫模型HMM，判别式分析，贝叶斯网络，Sigmoid Belief Networks，马尔科夫随机场，深度信念网络DBN

类别不平衡

1. 欠采样（undersampling）：也称下采样(downsampling)。去除一些反例样本使得正、反数据接近。然后再进行学习。

2. 过采样（重采样）（oversampling）：也称上采样(upsampling)。增加一些正例样本使得正、反数据接近。然后再进行学习。

3. 代价敏感（cost-sensitive learning）：增加将正例样本错判为负例样本的惩罚。将式（3）中的 m^+/m^-用cost^+/cost^- 代替即可，后面分别是正例误分为反例的代价，反例误分为正例的代价。

4. 阈值移动（threshold-moving）：直接基于原始训练集进行学习，但在用训练好的分类器进行预测时，将式（3）嵌入到其决策过程中。

决策树

**什么时候会产生叶子节点？**

1.当划分到训练集被划分到只有一类样本

2.当划分到没有属性可用于划分

3.当训练集在所有属性上取值一样

4.当划分到某个节点训练集为空——类别标记为不划分时对应的训练集中类别最多的点

**划分属性的选择算法**

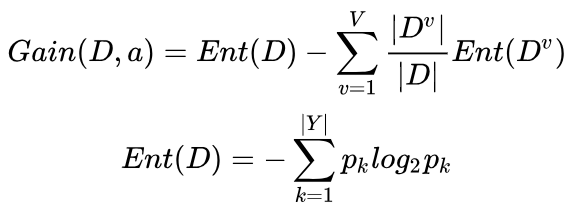
信息增益（ID3）

增益率（C4.5）

基尼指数（CART）

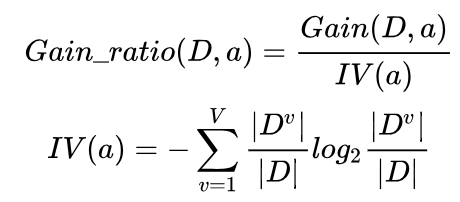
信息熵描述系统的混乱程度，值越大，越混乱，纯度越低

信息增益越大，使用该属性进行划分所获得的纯度提升越大



信息增益准则对可取值数目较多的属性有所偏好

增益率



IV(a)称为属性a的“固有值”（intrinsic value）。属性a的取值数目越多（即V越大），则IV(a)的值通常会越大。

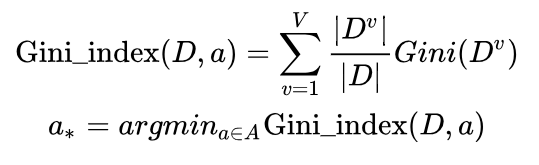
增益率对可取值数目较少的属性有所偏好。

C4.5算法并不是直接选择增益率最大的候选划分属性，而是使用了一个启发式：先从候选划分属性中找出信息增益高于平均水平的属性，再从中选择增益率最高的。

基尼指数



Gini(D)反映了从数据集D中随机抽取两个样本，其类别标记不一致的概率。因此Gini(D)越小，则数据集D的纯度越高。在候选属性集合A中，选择那个使得划分后基尼指数最小的属性作为最优划分属性



ID3和C4.5都是用于回归，CART既可以用于回归，也可以用于分类。

ID3不能处理连续属性

C4.5算法采用的是二分法，将连续属性a在数据集D中出现的取值情况进行排序，每相邻两个取值取其中间点作为候选划分点，基于每个候选划分点计算对应的信息增益，然后选择信息增益最大的对应的划分点。（也可将划分点设为该属性在训练集中出现的不大于中位点的最大值，从而使得最终决策树使用的划分点都在训练集中出现过）

CART既可以是分类树，又可以是回归树，而且它是一棵二叉树，也就是说每次只有两个分支。普通的决策树并没有这种限制，比如属性纹理在划分时有清晰、稍糊、模糊三个分支，到CART里就必须是类似于清晰、非清晰这样两个分支的情况。

CART对于连续属性的处理与C4.5类似，是这样的：

将所有属性j的取值拿出来（去掉重复值），进行排序，每个取值s都可以作为候选划分点。要么大于s，要么小于s。

对于离散属性，每个属性的取值s都可以作为候选划分点，要么等于s，要么不等于s。

当CART是分类树时，采用Gini指数作为节点划分的依据；当CART是回归树时，采用样本的最小方差作为节点划分的依据。

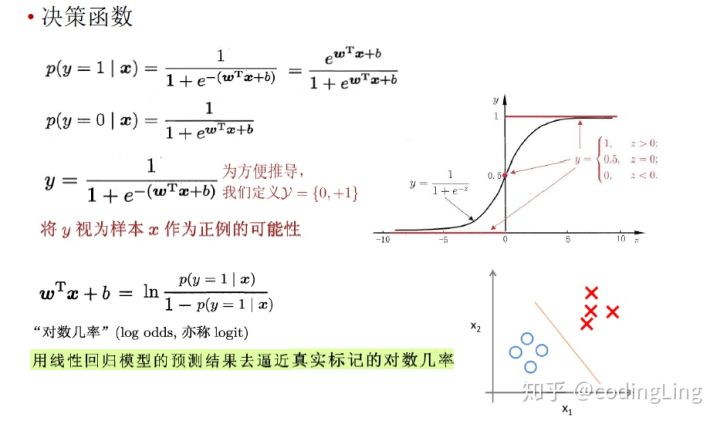
**剪枝**

预剪枝使得决策树的很多分支都没有“展开”，这不仅降低了过拟合风险，还显著减少了决策树的训练时间开销和测试时间开销。但另一方面，也有欠拟合的风险（有些分支的当前划分虽不能提高泛化性能，甚至可能导致泛化性能暂时下降，但在此基础上的后续划分可能导致泛化性能显著提高，但却被预剪枝“贪心”地剪掉了）。

后剪枝决策树通常比预剪枝决策树保留了更多的分支。一般情形下，后剪枝决策树的欠拟合风险很小，泛化性能往往优于预剪枝决策树，但后剪枝过程是在生成完全决策树之后进行的，并且要自底向上地对树中的所有非叶结点进行逐一考察，因此其训练时间开销比未剪枝决策树和预剪枝决策树都要大得多。

采用留出法，即预留一部分数据用作验证集

逻辑回归



对数几率回归方法的优点：

1. 它是直接对分类可能性进行建模，无需实现假设数据分布，这样就避免了假设分布不准确所带来的问题；

2. 它不是仅预测出“类别”，而是可得到近似概率预测，这对许多需利用概率辅助决策的任务很有用；

3. 对数函数是任意阶可导的凸函数，有很好的数学性质，现有的许多数值优化算法都可直接用于求取最优解。

LDA线性判别分析

将所有样本投影到一条远点出发的直线，使得同类样本距离尽可能小，不同类样本距离尽可能大，具体为最大化“广义瑞利商”。

SVM支持向量机

支持向量机与对率回归的优化目标相近，通常情形下它们的性能也相当.对率回归的优势主要在于其输出具有自然的概率意义，即在给出预测标记的同时也给出了概率，而支持向量机的输出不具有概率意义，欲得到概率输出需进行特殊处理[Platt ， 2000]; 此外，对率回归能直接用于多分类任务，支持向量机为此则需进行推广[Hsu and Li，2002]. 另一方面， hinge 损失有一块"平坦"的零区域，这使得支持向量机的解具有稀疏性，而对率损失是光滑的单调递减函数，不能导出类似支持向量的概念，因此对率回归的解依赖于更多的训练样本， 其预测开销更大.

SVR

对样本(x,y)，传统回归模型通常直接基于模型输出f(x) 与真实输出y 之间的差别来计算损失，当且仅当f(x) 与y 完全相同时，损失才为零.与此不同，支持向量回归(Support Vector Regression，简称SVR)假设我们能容忍f(x) 与y 之间最多有 \epsilon 的偏差，即仅当f(x) 与y 之间的差别绝对值大于 \epsilon 时才计算损失.

SVDD算法(Support Vector Data Description)

学习出一个最小的超球面（超球面是指3维以上的空间中的球面，对应的2维空间中就是曲线，3维空间中就是球面，3维以上的称为超球面），把这堆数据尽可能全都包起来，识别一个新的数据点时，如果这个数据点落在超球面内，就是这个类，否则不是。

集成学习（Stacking,Bagging,Boosting,Blending,RF,AdaBoost,GBDT,xgboost）

聚类（Kmeans,DBSCAN,层次聚类AGNES和DIANA）

特征工程

半监督学习

EM算法

GMM模型

最大熵模型

概率图模型

排序算法

二叉树

表示学习

推荐算法

CNN，RNN，LSTM等

简历

自我介绍