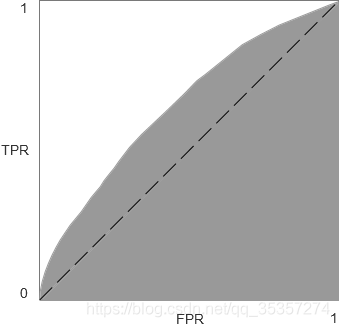
1. AUC与lrank
2. AUC

由这两个由这两个数据作为ROC曲线的横轴与纵轴，而曲线下方的面积为AUC。

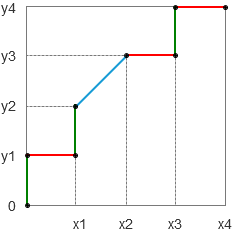


在FPR增大的情况下，TPR的值也会增加。原因是与分类的阈值设定有关。当阈值较小时，分类器对于正例预测的较为严格，所以反例被预测为正例的可能性较低，正例被预测为正例的可能性也比较低；而当阈值较大时，则情况相反。

预测结果：

(s1, 0.77, +),(s2, 0.62, −),(s3, 0.58, +),(s4, 0.47, +),(s5, 0.47, −),(s6, 0.33, −),(s7, 0.23, +),(s8, 0.15, −)

其中有4个正例，4个反例。

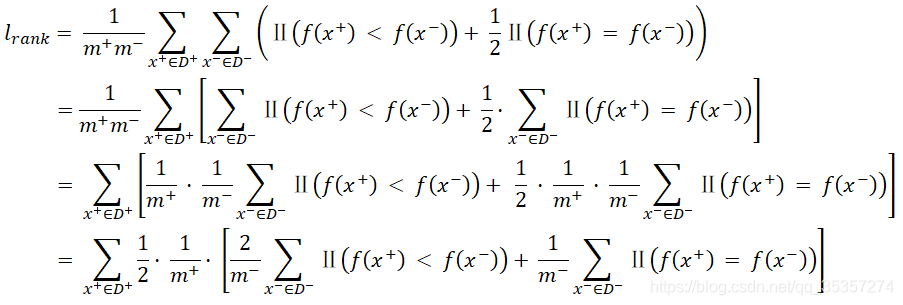


计算方法：

1. 将分类阈值设为最大，TPR和FPR都为0，第一个点(0, 0)；
2. 将分类阈值依次设为每个样例的预测值，即依次将每个样例划分为正例；
3. 前一个标记点坐标记为（x， y）；
4. 若当前为真正例则对应的标记点坐标为（x, y + 1 / m+），若当前为假正例则对应的标记点坐标为（x + 1 / m-, y）；
5. 相连节点（如果预测值相同，那么跳过中间那个节点，连接相同节点的前一个节点和最后一个相同的节点，比如上面数据中的s3与s5）。
6. Lrank

AUC考虑的是样本预测的排序质量，因此它与排序误差有密切的联系。

根据《南瓜书》上面的推导



这个公式意思为：当遍历所有正例的时候，排在它前面的有哪些反例比它大。相当于错误分类的概率。我们希望的理想分类情况时所有的正例排在反例之前。

1. 决策树
2. 熵的定义：

熵是一个事物的不确定性。

1. 熵如何进行量化

参照一个不确定事件作为单位（以抛硬币的不确定性为基本单位，单位是1bit）

一个事物的不确定性 = 抛几次硬币的不确定性

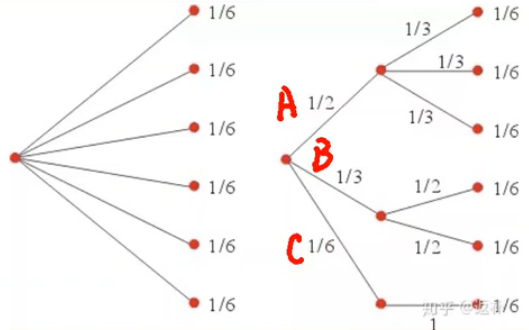
例如：

8个等概率事件的不确定性相当于抛3个硬币，熵为3bit。

1. 信息熵

的值越小，代表事件的不确定越低，纯度越高。

当事件发生的概率为不等概率事件时，就无法利用8个等概率事件那样去寻找熵。



先将事件划分为等概率事件，所以相当于个硬币的熵。之后由于A事件的三个分支是一个事件。所以会减少。所以A事件的熵为.同理可以推出B、C事件的熵值。所以整个系统的熵值为A事件的熵乘以A事件发生的概率。

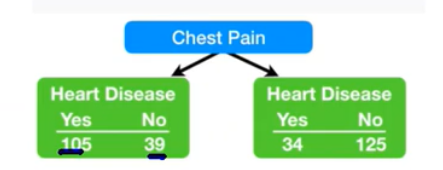
1. 信息增益

信息增益是假设用样本中的属性去划分样本，再划分样本之后。计算划分后整个系统的信息熵。之后与上一节点的信息熵相减。信息增益越大代表着通过这个属性值划分样本之后，整个系统的不确定性降低，纯度增加。

1. 基尼指数

基尼指数是另一种评判系统纯度的指标。

直观理解为，从化分好的数据集D中随机抽取两个样本，其类别不一致的概率。值越小代表着数据集纯度越高。



所以这个系统的基尼指数为：

[1 - ] + [1 - ]

根据基尼指数进行数据集的划分时，基尼指数应该从根节点向叶节点递减，如果是递增或者是不变则不应该设置该节点。

1. 剪枝处理

剪枝是决策树中处理“过拟合”的主要手段。分为“预剪枝”和“后剪枝”。

预剪枝是指在决策树生成过程中，对每个节点再划分前进行估计，若当前结点的划分不能提升决策树在测试集上的准确率则不划分；若可以提升在测试集上的准确率则划分。

后剪枝是指在决策树根据训练集建立完毕后，自底向上地对非叶节点进行考察，若将该节点对应的子树替换为叶节点能提升在测试集上的准确率则进行替换，否则不进行替换。

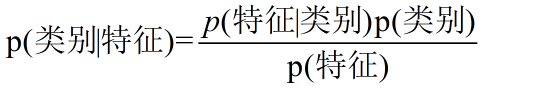
1. 贝叶斯分类器
2. 贝叶斯决策理论

每次在做判断时，选择选择后验概率最大的结果，就可以最小化预测的错误概率。

参考：https://www.zhihu.com/question/27670909

1. 朴素贝叶斯分类器

朴素贝叶斯算法是假设各个特征之间相互独立。



1. 极大似然估计

令表示数据集D中第C类样本组成的集合，假设这些样本都是独立同分布的，则参数对于数据集D的似然是

对进行极大似然估计，就是去寻找能最大化似然的参数值。极大似然估计是试图在所有可能的取值中，找到一个能使数据出现的“可能性”最大的值。

1. 似然函数

输入有两个：x表示某一个具体的数据；θ表示模型的参数。

如果θ是已知确定的，x是变量，这个函数叫做概率函数(probability function)，它描述对于不同的样本点x，其出现概率是多少。

如果x是已知确定的，θ是变量，这个函数叫做似然函数(likelihood function), 它描述对于不同的模型参数，出现x这个样本点的概率是多少。

1. 最大似然估计

参考：<https://zhuanlan.zhihu.com/p/26614750>

多数情况下我们是根据已知条件来推算结果，而最大似然估计是已经知道了结果，然后寻求使该结果出现的可能性最大的条件，以此作为估计值。

概率是根据条件推测结果，而似然则是根据结果反推条件。在这种意义上，似然函数可以理解为条件概率的逆反。

1. 最大后验概率
2. EM算法
3. 参考：<https://blog.csdn.net/v_july_v/article/details/81708386>
4. 拉普拉斯修正

为了避免其他属性携带的信息被训练集中未出现的属性值“抹去”，在估计概率值时通常要进行“平滑”，常用拉普拉斯修正。具体来说，令N表示训练集D中可能的类别数，表示第i个属性可能的取值数

1. 集成学习 --- 如何产生“好而不同”的个体学习器

集成学习通过将多个学习器进行结合，通常可以获得比单一学习器显著优越的泛化性能。

1. 为什么集成学习可以提高模型的性能

假设只有一个模型，整体的准确率是51.5%，那如果有三个模型进行集成，整体准确率就为： = 51.5%。那么假设有500个子模型，则整体的准确率是

1. Adaboost

参考资料：<https://zhuanlan.zhihu.com/p/27126737>

<https://blog.csdn.net/guyuealian/article/details/70995333>

1. GBDT

可以用来分类、回归和筛选特征。GBDT采用加法模型（即基函数的线性组合），以及不断减少训练过程产生的残差来达到将数据分类或者回归的算法。利用GBDT（决策树组成的森林）进行特征的选取

1. Bagging与随机森林

目的是为了保证“好而不同”，从训练集中进行抽样，组成每个基模型所需要的子训练集对所有基模型预测的结果产生最终的预测结果。Bagging通常对分类任务使用简单投票法，对回归任务使用简单平均法。若分类预测时出现两个类收到同样票数的情况，则最简单的做法是随机选择一个。

取样的方法有：

·Bagging:放回取样

·Pasting:不放回取样

·boostrap\_feature:在特征空间中随机采样

·即针对样本，又针对特征进行随机采样

1. 聚类
2. 聚类的评价指标
3. 外部指标： --- 选择一个“参考模型”进行比较，有点类似混淆矩阵

Jaccard系数、FM系数、Rand指数

三个指标越大说明聚类效果越好。

1. 内部指标

DB指数（值越小越好）、DI指数（值越大越好）

1. 聚类算法
2. K均值算法
3. 学习向量算法（LVQ）

试图找到一组原型向量来刻画聚类结构，但与一般聚类算法不同的是，LVQ假设数据样本带有类别标签，学习过程中利用样本的这些监督信息来辅助聚类。

1. 密度聚类

-领域：

核心对象：

密度直达：

密度可达：

密度相连：

基于这些概念，DBSCAN将“簇”定义为：由密度可达关系导出的最大的密度相连样本集合。

1. 层次聚类

参考：<https://zhuanlan.zhihu.com/p/50113029>

1. 高斯混合聚类
2. 降维
3. PCA降维

参考：<https://zhuanlan.zhihu.com/p/26951643>

在多元统计分析中，主成分分析（Principal components analysis，PCA）是一种分析、简化数据集的技术。主成分分析经常用于减少数据集的维数，同时保持数据集中的对方差贡献最大的特征。这是通过保留低阶主成分，忽略高阶主成分做到的。这样低阶成分往往能够保留住数据的最重要方面。

高维特征中很多特征之间存在相关性，含有冗余信息；相比于低维数据，高维数据计算更复杂。

1. 特征选择

进行特征选取之后会减少维数灾难的发生，从数据中选择出重要的特征；并且可以去除不相关特征，降低学习任务的难度。但是在特征选取的过程中要确保不丢失重要特征。所谓“无关特征”是指与当前学习任务无关的那些特征。所谓“冗余特征”是指包含的信息能从其他特征中推演出来。冗余特征在很多时候是不起作用的，去除它们会减轻学习的难度。

1. 贪心式进行特征选取

·前向搜索：从单一特征开始逐步增加特征。

·后向搜索：从完整的特征集合开始逐步删除特征。

·双向搜索：将前向搜索和后向搜索结合起来，在每一轮中逐渐增加选定相关特征（这些特征在后续轮中不会被去除）、同时减少无关特征。

1. 特征子集评价

可以使用“信息增益”、“信息熵”等指标进行评价。

1. 过滤式选择

过滤式的方法是先对数据集进行特征选择，然后再训练学习器，特征选择的过程与后续学习器无关。其中代表性的算法有Relief算法。

其中表示样本在属性j上的取值，取决于属性j的类型；若属性j为离散型，则时为0，否则为1；若属性j为连续型，则。

可以看出当这个量越大时，说明对应属性的分类能力就越强。

Relief算法是针对二分类设计的，它的扩展变体Relief-F能处理多分类问题。

1. 包裹式选择

与过滤式特征选择不考虑后续学习器不同，包裹式特征选择直接把最终将要使用的学习器的性能作为特征子集的评价准则。一般而言，由于包裹式特征选择方法直接针对给定学习器进行优化，因此从最终学习器性能来看，包裹式特征选择比过滤式特征选择更好，但是包裹式计算开销大的多。主要的算法有LVW。

1. 嵌入式选择与正则化

在过滤式和包裹式特征选择方法中，特征选择过程与学习器训练过程有明显的分别；与此不同，嵌入式特征选择是将特征选择过程与学习器训练过程融合在一体，两者在同一个优化过程中完成，即在学习器训练过程中自动地进行了特征选择。

主要选择的正则化方式有L1、L2正则化。