\*\*\*机器学习十大经典算法\*\*\*

1、Apriori（关联规则算法）

1.1 关联规则算法是数据挖掘的重要算法之一，主要有两个步骤：

1、**找出**交易数据库中所有大于或等于用户指定的最小支持度的**频繁项集**。

2、利用频繁项集生成所需要的关联规则，根据用户设定的最小可信度**筛选**出强**关联规则**。

1.2两个重要概念：支持度、可信度，以此来度量关联规则的强度

支持度：确定规则可以用于给定数据集的频繁程度；（用自己的话理解最好：多少人支持你）

置信度：确定Y在包含X的事务中出现的频繁程度。（某人的话有多少可信）

1.3 算法优缺点

优点：**简单易理解**；使用先验性质，大大提高了频繁项集逐层产生的效率；数据集要求低。

缺点：1、候选频繁K**项集数量巨大**；2、在验证候选频繁K项集的时候，需要对整个数据库进行扫描，非常耗时，**效率低下**。

1.4 改进算法

2、Kmeans（K平均算法）

1.1 最著名的聚类方法，基本原理：

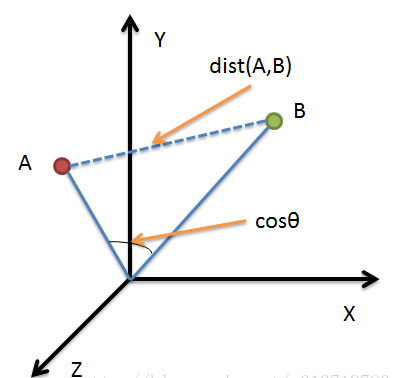
重复移动类的中心点（重心)，移动到其包含成员的平均位置，然后重新划分其内部成员，成本函数是该类重心与其内部成员位置距离的平方和，最优解是成本函数最小时k的取值。

k是算法表示类的数量，必须是一个比训练集样本数小的正整数。

Kmeans可以自动分配样本到不同的类，但是不能决定究竟要分几个类。

1.2 欧式距离与余弦相似度

欧氏距离用于计算两点距离度量：体现个体数值特征的**绝对差异**，所以更多的用于需要从维度的数值大小中体现差异的分析，如使用用户行为指标分析用户价值的相似度或差异；



余弦相似度则是最常见的相似度度量：**从方向上区分差异，对绝对的数值不敏感**，更多的用于使用用户对内容评分来区分用户兴趣的相似度和差异，同时修正了用户间可能存在的度量标准不统一的问题。

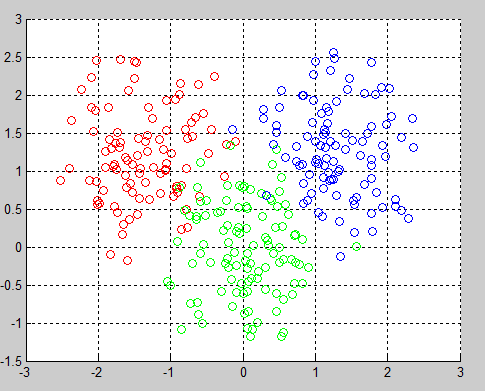
1.3 优缺点

优点：算法**原理简单**（需要调节的超参数就是一个k），**效率高**和可扩展性良好；

缺点：1、**k值难确定**，且选择的好坏直接影响能否得到有效聚类结果；2、当数据量非常大时，算法的时间开销是非常大的；3、基于很多理想假设（A、各个 cluster 的先验概率相同，B、数据的各个维度对于相似度的衡量作用相同），但这些假设现实数据中很难满足。

1.4 效果示意

当k取3时，聚成3类



3、SVM（支持向量机）

4、Logistic Regression（逻辑回归）

5、EM（最大似然估计）

6、KNN（K最近邻）

1.1 KNN是一种用于分类和回归的非参数统计方法，其核心思想是：

如果一个样本的K个最相邻样本中的大多数属于某一个类别，则该样本也属这个类别，并具有该类别上样本的特性。

1.2 K的选择

如何选择K值：一般情况下，K值较大可以减少噪声影响，但同时会模糊类别界限，一个较好K值可通过启发式技术获取。计算样本距离采用欧式距离。

1.3 优缺点

优点：1、简单易理解，易实现；2、无需训练；3、特别适合于多分类问题。

缺点：1、好的K值的方法只能是通过反复试验调整；2、随着样本N增加算法复杂度较高O(N)；

1.4 应用场景

首选给出多部电影中的各类镜头（搞笑、打斗、拥抱）数量及类型，计算一个新上映的电影所属的类型（喜剧、动作、爱情片）。



7、Native Bayes（朴素贝叶斯）

1.1 朴素贝叶斯模型与决策树模型为应用**最广泛的两个**分类模型。其核心思想

选择条件概率较大的类别作为分类（决策）的结果。

1.2 重要概念

**先验概率**：【经验概率】个人觉得更形象。也就是说根据过往经验和数据计算得出的属于某一个分类的概率。与之相对的是后验概率，也就是加入其他附加信息后计算得到概率值。如根据今天及之前的天气情况推断明天下雨的概率就是先验概率，如果再加上卫星云图等信息预测明天下雨概率则为后验概率。

**贝叶斯定理**：描述两个条件概率之间的关系，P(A∩B) = P(A)\*P(B|A)=P(B)\*P(A|B)。统计学的基本理论，现实意义可翻译为“支持某项属性的事件发生得愈多，则该属性成立的可能性就愈大”。

朴素贝叶斯：**朴素**就是简单的意思，之所以简单，是因为其假设【给定目标值时各属性之间相互独立】，贝叶斯公式中的A、B就是不同的维度。

1.3 优缺点

优点： 1、 算法逻辑简单,易于实现（使用贝叶斯公式转即可！）；2、算法复杂度很小（假设特征相互独立，只会涉及到二维存储）

缺点：在属性个数比较多或者属性之间相关性较大时，分类效果不好。有半朴素贝叶斯之类的算法通过考虑部分关联性适度改进。

1.4 示例场景

1、足球比赛中，判断A队与B队最终是否会在决赛中相遇。

2、一个姑娘相亲，根据男方多方面条件，决定是要跟他确定关系。

8、C4.5（决策树）

1.1决策树模型与朴素贝叶斯模型为**两个最经典的**分类模型。改进之后的C4.5决策树算法既适合于分类问题，又适合于回归问题。决策树分类的基本步骤如下：

1、生成决策树：使用样本数据，利用归纳算法生成可读的规则和决策树；

2、决策树剪枝：使用新的数据校验决策树中的初步规则，进行检验、校正和完善决策树；

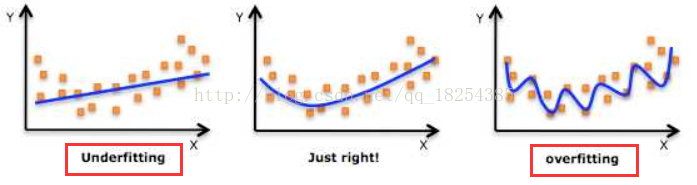
1.2 优缺点

优点： 1、**简单直观**，算法代价是O(log2m)很低。2、 可以处理多维度输出的分类问题；既可以处理离散值或**连续值**。 3、 对于异常点的容错能力好，**健壮性高**。

缺点：1、决策树算法非常容易过拟合，导致泛化能力不强；2、不稳定：决策树会因为样本发生一点点的改动，就会导致树结构的剧烈改变；3、样本欺骗：某些特征的样本比例过大，生成决策树容易偏向于这些特征；

1.3 重要概念

**拟合**：形象的说，拟合就是把平面上一系列的点，用一条光滑的曲线连接起来。因为这条曲线有无数种可能，从而有各种拟合方法。



过拟合（over-fitting）是指模型在**训练样本**中表现过于**优越**，导致在验证数据集以及**测试数据集**中表现**不佳**。【原因】造成过拟合的原因可以归结为：参数过多，学习了数据中的噪声。【解决措施】在训练数据的子集上训练和测试模型k次；保留一个验证数据集；是减少参数，对于神经网络可采用 dropout方法，其的思想是当一组参数经过某一层神经元的时候，去掉这一层上的一部分神经元，让参数只经过一部分神经元进行计算。

欠拟合（Underfitting）：是指通过样本训练得到模型效果较差，未达到预期效果。【原因】欠拟合一般发生在训练刚开始的时候【解决措施】增加网络复杂度或者在模型中增加多个特征点或维度可消解。

举例来说：把高中生当做一个模型，模拟考试发挥正常，高考发挥失常就是过拟合，虽然在训练集上表现得很好，但是在测试集中表现得恰好相反。平时模拟考试就发挥不正常，更别提高考了，这就是欠拟合。

所谓欠拟合呢（under-fitting）？相对过拟合欠拟合还是比较容易理解。还是拿刚才的模型来说，可能二哈被提取的特征比较少，导致训练出来的模型不能很好地匹配，表现得很差，甚至二哈都无法识别

**泛化**：就是举一反三

9、CART（分类与回归树）

10、Random Forest（随机森林）

\*\*\*数据挖据四大问题\*\*\*

A、分类

B、聚类

C、关联

D、预测

这里我们对FP Tree算法流程做一个归纳。FP Tree算法包括三步：

　　1）扫描数据，得到所有频繁一项集的的计数。然后删除支持度低于阈值的项，将1项频繁集放入项头表，并按照支持度降序排列。

　　2）扫描数据，将读到的原始数据剔除非频繁1项集，并按照支持度降序排列。

　　3）读入排序后的数据集，插入FP树，插入时按照排序后的顺序，插入FP树中，排序靠前的节点是祖先节点，而靠后的是子孙节点。如果有共用的祖先，则对应的公用祖先节点计数加1。插入后，如果有新节点出现，则项头表对应的节点会通过节点链表链接上新节点。直到所有的数据都插入到FP树后，FP树的建立完成。

　　4）从项头表的底部项依次向上找到项头表项对应的条件模式基。从条件模式基递归挖掘得到项头表项项的频繁项集。

　　5）如果不限制频繁项集的项数，则返回步骤4所有的频繁项集，否则只返回满足项数要求的频繁项集。