数据挖掘&机器学习

分类：从机器学习的角度，分类属于监督学习，每个训练样本的数据对象

已经有类标识，通过学习可以形成表达数据对象与类标识间对应的知识。

问题（Problem）:

1) 噪声数据导致过度拟合

2) 缺乏代表性样本导致过度拟合

聚类：从机器学习的角度，聚类属于无监督学习，每个训练样本的数据对象是没有类别标识的。聚类的目的是使得属于同类别的对象之间的差别尽可能的小，不同类别的对象尽可能的大。

K-means：

聚类是一个将数据集中在某些方面相似的数据成员进行分类组织的过程，聚类就是一种发现这种内在结构的技术，聚类技术经常被称为无监督学习。K均值聚类算法是先随机选取K个对象作为初始的聚类中心，然后计算每个对象与各个种子聚类中心之间的距离，把每个对象分配给距离它最近的聚类中心。k-means聚类算法是局部收敛的；k-means的聚类结果和初始聚类中心点的选取有关；k-means聚类算法的时间复杂度是O(NKT)，其中N为样本个数，K为聚类团数量，T为迭代次数；k-means聚类算法无法自动确定聚类团数量。

K-Means的主要优点有：①原理比较简单，实现也是很容易，收敛速度快。②聚类效果较优。③算法的可解释度比较强。④主要需要调参的参数仅仅是簇数k。K-Means的主要缺点有：①K值的选取不好把握②对于不是凸的数据集比较难收敛③如果各隐含类别的数据不平衡，比如各隐含类别的数据量严重失衡，或者各隐含类别的方差不同，则聚类效果不佳。④采用迭代方法，得到的结果只是局部最优。⑤对噪音和异常点比较的敏感。

判别式模型（Discriminative Model）是直接对条件概率p(y|x;θ)建模。

常见的判别式模型有 线性回归模型、线性判别分析、支持向量机SVM、神经网络等。

生成式模型（Generative Model）则会对x和y的联合分布p(x,y)建模，然后通过贝叶斯公式来求得p(yi|x)，然后选取使得p(yi|x)最大的yi，

常见的生成式模型有 隐马尔可夫模型HMM、朴素贝叶斯模型、高斯混合模型GMM、LDA等。

分类器集成，其实就是集成学习，通过构建并结合多个学习器来完成学习任务。一般结构是：先产生一组“个体学习器”，再用某种策略将它们结合起来。结合策略主要有平均法、投票法和学习法等。这些策略使得集成提高了性能。