第一：分类和聚类的区别

通过学习得到一个目标函数，把属性集x映射到一个先定义好的类别中

**分类**是根据已经给定的类别标号样本，得到目标函数，从而能够对未知属性的类别的样本进行分类，目的找到样本属性和类别之间的关系

**聚类**：事先不知道任何样本的标号，希望通过算法来把未知类别的样本划分成若干类别。聚类目的是将相似的东西放到一起，一般被称作无监督学习。

目标是组内相似性越大，而组间差别越大。

**半监督学习**：极少量的有类别标签样本，和大量的无类别标签样本

分类中的生成方法和判别方法区别：

**生成方法**：学习联合概率分布P(X,Y)，然后求出条件概率分布P(Y|X)作为预测的模型，可以简化学习问题，需要更多的样本

**判别方法**：直接学习决策树函数f(X)，或者条件概率分布P(Y|X)作为预测模型

可以直接用于预测，往往学习的准确率较高

分类方法总结

1. **贝叶斯**

基于贝叶斯定理的统计学分类方法，通过预测一定给定元素属于某个类的概率来进行分类

* **朴素贝叶斯**：假定不同属性的相互独立

优点：所需参数少，对数据缺失不敏感

缺点：假设往往不成立，需要知道先验概率

**贝叶斯网络**

1. **决策树**

每个节点表示在某个属性上的测试，不同的分支表示不同的测试结果

基于统计方法构建决策树

ID3：贪心方法，保证当前的节点是最优节点

优点：无参数假设、适合高维数据，易于理解、短时间内可以处理大量数据

缺点：易于过拟合，忽略了属性间的相关性，不支持在线学习

信息熵、信息增益、分裂信息、增益比

基尼系数

CART：决策树是二叉树，最小平方误差准则

1. **SVM支持向量机**

把分类问题转换为寻找分类平面问题，并通过最大化分类边界点距离分类平面来实现分类，对偶问题

优点：小样本学习、泛化性高、高维、非线性

缺点：缺失数据敏感

1. **KNN K-近邻**

用最近的K个点的类别表示当前点的类别

缺点：计算量大，样本分类不均

1. **LR 逻辑回归**

优点：简单速度快，易于理解

缺点：处理复杂，需要归一化和较多的特征工程

1. **神经网络**

梯度下降法

1. **集成学习**

**GBDT：**

Gradient Boosting + 决策树 = GBDT

**Adaboost**

通过训练数据学习一系列弱分类器，然后由这些弱分类器组合成一个强分类器

学习，降低正确样本权值，增加错误样本权值，每次训练改变样本权值

组合：加权多数表决

**Bagging：**

可以并行进行

从原始样本中随机均匀抽取k个训练集，每个训练集n个训练样本，学习k个模型，k个模型进行加权表决。

**Random Forest**

多颗决策树联合组成的预测模型，可以对样本或特征值取bagging

Bagging + 决策树 = 随机森林

**Boosting**

聚类算法

**K-means**，K-media，K-

1、初始选择K个类别中心，2、将每个样本标记为距离类别样本中心最近的那个类别，3、将每个类别中心更新为这个类别中所有点的均值，4、重复2、3

缺点：异常点会导致中心的偏离，无法处理奇怪形状，K值无法确定

优点：简单、快速，易于理解

**DBSCAN**

有两个参数：最小数目m，半径sigma。将点分为3类，核心点，边缘点，噪声点

核心点：核心对象的半径sigma内至少有m个点。

边缘点：在核心点的半径sigma内的点

优点：没有形状限制，可以剔除噪声点。

缺点：sigma和m的值的确定上，同时，只能对同密度的聚类

**DPEAK算法**

给一个半径r，计算每个样本r邻域内的密度为rho，样本密度比它高的点的距离的最小值记作sigma

1. rho小，但sigma很大，周围样本少，距离中心远，异常值
2. rho很大，sigma很大，周围样本多，距离中心远，簇中心
3. rho很小，sigma很小，周围样本少，距离中心近，边缘点
4. rho很大，sigma很大，周围样本多，距离中心近，核心

优点：确定簇中心和排除异常值

**OPTICS算法**