■数据预处理:对原始数据进行清洗、集成、变换、维度规约、数值规约

数据库与数据库的区别

外部表与内部表的区别

■OLAP 依赖于如 Hive 这样的数据仓库而非数据库.
■为什么不能直接在普通数据库上进行操作而要使用数据仓库
.■提高系统的性能:数据库是为 OLTP 而设计的，数据仓库是为 OLAP 而设计，为复杂的 OLAP 查询，多维视图，汇总等 OLAP 功能提供了优化。，
功能不同:数据库支持多事务的并行处理，而数据仓库往往只是对数据记录进行只读访问
■数据不同:数据仓库中存放历史数据;日常数据库中存放的往往为最新数据。

机器学习算法的分类
有监督(指导)学习(supervised learning)
■有训练数据集
■回归(regression)
■线性回归，如最小二乘法
■线性回归旨在拟合一条曲线表达式，到达所有样本点的距离的和是最小的
■分类(classification)
■决策树
■SVM(support vector machine)
■分类的输出为离散的类别，而回归可以计算连续数值。
■无监督(指导)学习(unsupervised learning)
■无训练数据集
■聚类，如 k-means .

分类算法
SVM (Support Vector Machine) 算法的优缺点
■优点:小数据集，使用不同的核函数可以支持非线性分类，通过调整参数可以获得高精度分类，泛化能力好
缺点:运算太慢
■数据干净、线性可分有利于 SVM 算法，有噪声会导致 SVM 算法性能下降
■训练好后，抛弃非支持向量的样本点，仍然可以对新样本进行分类。
分割平面由 supportvector 决定

找到最佳的决策树是 NP 问题。
即使训练集很大，构建决策树的代价也较小。
决策树相对容易解释。
决策树算法对于噪声的干扰具有相当好的鲁棒性。
冗余属性不会对决策树的准确率造成不利的影响。
子树可能在决策树中重复多次，使得决策树过于复杂，并且更难解释。

■信息论中信息量与事件概率的关系

和关联规则挖掘过程是发现满足最小置信度的所有项集(item set)代表的规则

■置信度和支持度是不同的

■Apriori 先验原理:如果一个项集是频繁的，那包含它的所有非空子项集也是频繁的

，如果规则 A→C-A 不满足置信度阈值，则形如 A'→C-A'的规则有可能满足置信度要求，其中 A'是 A 的子集;

■油条、大饼->豆浆

■油条->豆浆、大饼