第一章 数据挖掘导论

一、为什么数据挖掘

数据丰富，但是信息匮乏，我们被海量的数据淹没，但是却缺少知识，针对这么多数据不知道怎么进行利用

在大数据时代，**我们虽然拥有丰富的数据，但是却缺乏有用的信息**，我们要将**数据提升到知识层面来支持决策**

二、什么是数据挖掘

1. CRISP-DM是一个迭代式的，具有自适应性的过程，这是跨行业数据挖掘的标准流程
2. 数据挖掘的概念
   - 数据挖掘是对大量的数据进行探索和分析，来获取有意义的模式
   - 数据挖掘是一项复杂的任务，其目的在于从海量的数据中其目的在于从海量数据中识别出有效、新颖、可能具有实用价值且最终能够被理解的模式。
   - **数据挖掘是从海量数据中获取正确的、新颖的、潜在有用的、最终可理解的模式的非凡过程**
   - **数据挖掘是从大量的、不完全的、有噪声的、模糊的数据中挖掘那些有用的、隐含的、先前未知的模式或知识的过程**
   - **数据挖掘是一个跨学科学习，数据挖掘来源于众多学科领域包括统计学，机器学习，人工智能和数据库管理。这种多学科交叉的方法使得数据分析师能够用各种各样的技巧和工具去有效解决不同的数据分析难题**

三、数据挖掘活动

1. **数据挖掘算法分类**

有监督学习	无监督学习
分类算法 决策树 KNN 朴素贝叶斯（NB） SVM 人工神经网络（ANN） **回归算法** 线性回归 多项式回归	**聚类算法** K-means 密度聚类 层次聚类 … **关联分析** Apriori FP-Growth Eclat …

- **分类：根据训练数据集和类标号属性，构建模型来分类现有数据，并且来分类新数据或者用来预测类型标志未知的对象类。**
- **聚类：按照某个特定标准将一个数据集分割成不同的类，使得类内相似性尽可能地大，同时类间的区别性也尽可能地大。**
- **关联分析：从大量数据集之间发现有用的、频繁出现的模式、关联和相关性。**
- **回归：通过对已有的数据进行归纳、学习，从而得到相应模型，再将该模型用于未知变量的预测。**
- 异常值发现：从正常数据中检测异常点，分析其意义