**第一章**

1. **数据挖掘：**

数据挖掘是通过对(大规模)观测数据集的分析,寻找确信的关系,并将数据以一种可理解

的且利于使用的新颖方式概括数据的方法

1. **监督学习：**

需要训练数据集，通常用于预测，指定了我们想要预测的值的可能值，并且我们有一些其值已知的训练案例。对于分类问题求分类的类别，对于回归问题预测数值，为了预测需要大量的数值。

1. **无监督学习**

对于无监督的学习问题，我们不关注任何特定事物的预测，而是试图找到数据的有趣方面

聚类

1. **半监督学习**

**介于监督学习和半监督学习之间的技术，**与监督学习相比，目标是减少所需的监督量。同时将无监督聚类的结果提高到用户的期望（约束聚类）

1. **强化学习**

**从互动中学习;学习做什么及其效果；目标导向的学习；**rial-and-error搜索和延迟奖励

1. **主动学习**

传统的监督学习算法被动地接受训练数据。主动学习从未标记的数据中获得信息；主动学习可以大大减少训练规模！

不确定性抽样

1. **ML和DM 与传统统计学**

**动机：预测--理解--因果关系**

许多传统统计数据主要是通过表明一个因素导致另一个因素而产生的。接下来是理解，最后预测。这些方法通常假设线性关系

1.经典（频率论）和贝叶斯哲学竞争。  2.不愿意使用没有理论依据的方法

3.通过情节和诊断来强调使用人类判断。

4.基于科学知识的模型。

5.最初设计用于手动计算，但计算现在非常重要

在机器学习和数据挖掘中，顺序通常是相反的：预测是最重要的。这些问题有很多变数

1.没有解决哲学或广泛接受的理论框架。

2.如果它们似乎运作良好，

3.强调自动方法，很少或没有人为干预。

4.适用于许多问题的方法。

5.大量使用计算

1. ML的挑战
2. 处理复杂性
3. 优化和整合（找到最佳参数）
4. 可视化
5. 维数爆炸

**限制复杂性，降维，平均复杂性**

# 第二章

## 贝叶斯理论

 Bayesian决策理论是数据挖掘/模式识别的统计方法 决策的数学基础 采用概率方法帮助做出决策，以最大限度地降低风险（成本）

基本假设:

决策问题是概率术语的参数形式，所有相关概率值已知