1. **无监督学习概论**

无监督学习：从无标注数据中学习模型

基本想法：对给定的数据(矩阵数据)进行某种**”压缩”**，从而找到数据的潜在结构、

**硬聚类(hard clustering):**一个样本只能属于一个类

**软聚类(soft clustering)**:一个样本可以属于多个类

硬聚类输出的是一个值，而软聚类输出的则是一个分布

聚类可以帮助发现数据中隐藏的纵向结构

**降维(dimensionality reduction)**:将训练数据中的样本从高维空间转换到低维空间

降维的模型是一个函数，输入高维的向量，输出低维的向量，可以是线性的也可以是非线性的

降维可以帮助发现数据中的横向结构

**概率模型估计(probability model estimation)**：由训练集学习概率模型的结构和参数

**无监督学习三要素**：模型、策略、算法

**模型**：函数z=gθ(x)或者条件概率分布Pθ(z|x)

**策略**：都可以表示为目标函数的优化

**算法**：通常是迭代算法

1. **聚类方法**

根据样本的特征相似度/距离，将它们归到若干个类里

**14.1 距离的基本概念**

14.1.1 相似度或距离

样本集合可以用矩阵X表示，矩阵元素xij表示第j个样本的第i个属性

可以用于衡量的距离：**闵可夫斯基距离，马哈拉诺比斯距离，相关系数，夹角余弦**

选取不同的相似度度量可以得到不同的结果

14.1.2 类或簇

定义：**类的均值、类的直径，散布矩阵，样本协方差矩阵**

散布矩阵看定义像是描述数据的离散程度

14.1.3 类和类之间的距离

类和类之间的距离又称为**链接(linkage)**

**最短距离或单链接、最长距离或完全链接、中心距离、平均距离**