第13章 无监督学习概论

1.机器学习或统计学习一般包括监督学习、无监督学习、强化学习。

无监督学习是指从无标注数据中学习模型的机器学习问题。无标注数据是自然得到的数据，模型表示数据的类别、转换或概率无监督学习的本质是学习数据中的统计规律或潜在结构，主要包括聚类、降维、概率估计。

2.无监督学习可以用于对已有数据的分析，也可以用于对未来数据的预测。学习得到的模型有函数$z=g(x)$，条件概率分布$P(z|x)$，或条件概率分布$P(x|z)$。

无监督学习的基本想法是对给定数据（矩阵数据）进行某种“压缩”，从而找到数据的潜在结构，假定损失最小的压缩得到的结果就是最本质的结构。可以考虑发掘数据的纵向结构，对应聚类。也可以考虑发掘数据的横向结构，对应降维。还可以同时考虑发掘数据的纵向与横向结构，对应概率模型估计。

3.聚类是将样本集合中相似的样本（实例）分配到相同的类，不相似的样本分配到不同的类。聚类分硬聚类和软聚类。聚类方法有层次聚类和k均值聚类。

4.降维是将样本集合中的样本（实例）从高维空间转换到低维空间。假设样本原本存在于低维空间，或近似地存在于低维空间，通过降维则可以更好地表示样本数据的结构，即更好地表示样本之间的关系。降维有线性降维和非线性降维，降维方法有主成分分析。

5.概率模型估计假设训练数据由一个概率模型生成，同时利用训练数据学习概率模型的结构和参数。概率模型包括混合模型、率图模型等。概率图模型又包括有向图模型和无向图模型。

6.话题分析是文本分析的一种技术。给定一个文本集合，话题分析旨在发现文本集合中每个文本的话题，而话题由单词的集合表示。话题分析方法有潜在语义分析、概率潜在语义分析和潜在狄利克雷分配。

7.图分析的目的是发掘隐藏在图中的统计规律或潜在结构。链接分析是图分析的一种，主要是发现有向图中的重要结点，包括 **PageRank**算法。

本文代码更新地址：https://github.com/fengdu78/lihang-code

中文注释制作：机器学习初学者公众号：ID:ai-start-com

配置环境：python 3.5+

代码全部测试通过。