读书报告

1. 别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题1：夹角余弦作为相似度是如何应用的

自己的理解：夹角余弦主要是判断方向的差异性，在文本分析中有相关的应用，讲文本中的词表示成向量形式进行计算可以用来比较文本之间的相似度。

1. 问题2：k均值据类属于启发式方法，什么叫做启发式的方法？

自己的理解：启发式的方法就是在给定的时间和空间内，给出问题的一个可行解，但是与最优解的偏差无法估计。

1. 读书计划

1、本周完成的内容章节：13、14

2、下周计划：15章节

1. 读书摘要
   * 1. 无监督学习是从无标注的数据中学习数据的统计规律或者说内在结构的机器学习，主要包括聚类、降维、概率估计。无监督学习可以用来数据分析或者监督学习的前处理。
     2. 无监督学习是一个困难的任务，因为数据没有标注，也就是没有人的指导，机器需要自己从数据中找出规律。模型的输入x在数据中可以观测，而输出z隐藏在数据中。无监督学习通常需要大量的数据，因为对数据隐藏的规律的发现需要足够的观测。无监督学习的基本思想是对给定数据进行某种“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。
     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。聚类时，样本通常是事先给定的。样本之间的相似度或距离由应用决定。如果一个样本只能属于一个类则称为硬聚类，如果一个样本可以属于多个类，则称为软聚类。
     4. 降维是将训练数据中的样本（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，或者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。
     5. 概率模型估计简称概率估计，假设训练数据由一个概率模型生成，由训练数据学习概率模型的结构和参数。
     6. 机器学习三要素：模型、策略、算法。
     7. 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。图分析：图数据表示实体之间的关系，包括有向图、无向图、超图。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。
     8. 相似度或距离：聚类的核心概念是相似度或距离，有多种相似度或距离的定义。因为相似度直接影响聚类的结果，所以相似度的选择是聚类的根本问题。闵可夫斯基距离：越大相似度越小，距离越小相似度越大。马哈拉诺比斯距离：距离越大相似度越小，距离越小相似度越大。相关系数：绝对值越接近于1，表示样本越相似，越接近于0，表示样本越不相似。夹角余弦。
     9. 类与类之间的距离称为连接，有最短距离或单连接，最长距离或完全连接、中心距离、平均距离。
     10. 层次聚类：层次聚类假设类别之间存在层次结构，将样本聚到层次化的类中。层次聚类又有聚合或下而上聚类、分裂或自上而下聚类两种方法。层次聚类是一种硬聚类
     11. 聚合聚类的三要素：距离或相似度；合并规则；停止条件。