方骏-2020年7月12日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：夹角余弦作为相似度是如何应用的

讨论后的理解：最常见的应用就是计算文本相似度。将两个文本根据他们词，建立两个向量，计算这两个向量的余弦值，就可以知道两个文本在统计学方法中他们的相似度情况。实践证明，这是一个非常有效的方法。

1. 别人提出的问题的理解：

2、 问题2：k均值聚类属于启发式方法，什么叫做启发式的方法？

自己的理解：一个问题的最优算法求得该问题每个实例的最优解。启发式算法可以这样定义：一个基于直观或经验构造的算法，在可接受的花费（指计算时间和空间）下给出待解决组合优化问题每一个实例的一个可行解，该可行解与最优解的偏离程度一般不能被预计。现阶段，启发式算法以仿自然体算法为主，主要有蚁群算法、模拟退火法、神经网络等。

1. 读书计划

1、本周完成的内容章节：《统计机器学习》第十四章

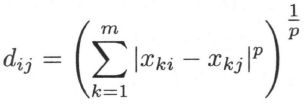
2、下周计划：《统计机器学习》第十五章

四、读书摘要及理解

1、聚类是针对给定的样本，依据它们特征的相似度或距离，将其归并到若干个"类"或"簇"的数据分析问题。一个类是样本的一个子集。直观上，相似的样本聚集在相同的类，不相似的样本分散在不同的类。这里，样本之间的相似度或距离起着重要作用。聚类算法很多，这里介绍的是两种最常用的聚类算法:层次聚类和k 均值聚类。

2、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。

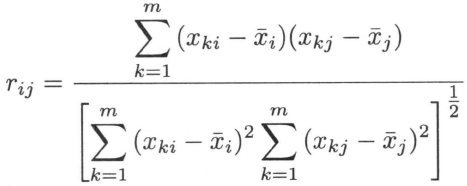
闵可夫斯基距离，用空间点之间的距离表示：



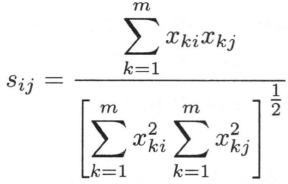
马哈拉诺比斯距离，考虑各个分量之间的相关性井与各个分量的尺度无关：



相关系数，样本之间的相似度也可以用相关系数C来表示：

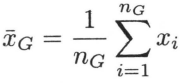


夹角余弦，样本之间的相似度也可以用夹角余弦来表示：



3、类的特征可以通过不同角度来刻画，常用的特征有下面三种：

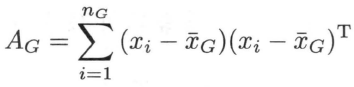
类的中心，即类的均值：



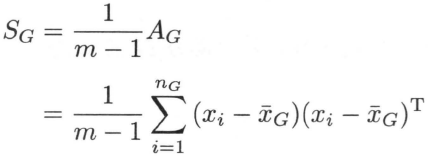
类的直径，类中任意两个样本之间的最大距离：



类的样本散布矩阵与样本协方差矩阵，类的样本散布矩阵：



样本协方差矩阵：



4、类与类之间的距离也称为连接，也有多种定义：

最短距离或单连接：



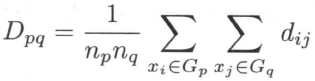
最长距离或完全连接：



中心距离：



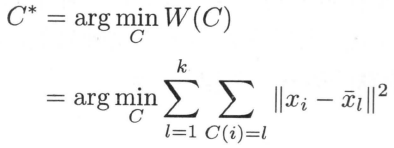
平均距离：



5、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自下而上聚类、分裂或自上而下聚类两种方法。聚合聚类开始将每个样本各自分到一个类:之后将相距最近的两类合井，建立一个新的类，重复此操作直到满足停止条件:得到层次化的类别。分裂聚类开始将所有样本分到一个类:之后将己有类中相距最远的样本分到两个新的类，重复此操作直到满足停止条件:得到层次化的类别。本书只介绍聚合聚类。

6、k 均值聚类是基于样本集合划分的聚类算法。k均值聚类将样本集合划分为k个子集，构成k个类，将n个样本分到k个类中，每个样本到其所属类的中心的距离最小。

k 均值聚类归结为样本集合X的划分，或者从样本到类的函数的选择问题。k均值聚类的策略是通过损失函数的最小化选取最优的划分或函数C\*。k 均值聚类就是求解最优化问题：



k 均值聚类的算法是-个迭代的过程，每次迭代包括两个步骤。首先选择k个类的中心，将样本逐个指派到与其最近的中心的类中，得到一个聚类结果；然后更新每个类的样本的均值，作为类的新的中心：重复以上步骤，直到收敛为止。