统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. 聚类和降维代表了从两种方向进行压缩的方法，那么这两种方法之间有什么相似性和差异？

讨论后的理解：

聚类考虑的是样本点间属性的相似性，这种相似性可以通过范数，相关性分析来衡量，因为属性在矩阵中代表的是行，所以被称为纵向压缩。而降维则是压缩属性的数量，属于横向的压缩。其目标是保留最重要的属性，体现出数据潜在的根本结构，但是在可视化后，可以看出有些降维后的数据产生了聚集，所以从某种程度上看这也是一种可以体现相关性的方法。

1. 聚类相似度的四种标准都是从什么角度考虑样本相似性的？

讨论后的理解：

从范数角度来看，其计算的是样本点间的距离指标，而三角函数和相关性的话，将夹角的大小作为指标，相关性还考虑了向量的均值。

1. 定义14.5如何推导定义14.8？

讨论后的理解：

可以考虑样本中所有的点，计算其和再求平均，然后根据14.5即可推导。

1. 别人提出的问题的理解
2. 降维降的维数由什么决定？

自己的理解:

降维的维数一般使用主成分分析中的方法来决定，一般将信息的保留量设定一个比例，然后通过计算可以计算出至少多少维度可以保留这么多的信息，然后就可以计算出维数。

1. 如何对无监督学习的结果进行评估？

自己的理解：无监督学习没有通用的评估方法，要根据具体的应用来评估，比如评价聚类的效果要看聚类后的聚类是不是高内聚的，有没有某些性质。对于降维要观察其能否保存足够的信息，在实际情况中是不是吻合等。

1. LDA是什么？

自己的理解：

LDA是潜在迪利克雷分配，是用于文本分析的一种概率模型，其是潜在语义分析和概率潜在语义分析的扩展。

1. 什么是无标注数据？他和有标注有数据的差异体现在哪里？

自己的理解:

无标注数据直观理解应该是没有人为进行标注过的数据，比如没有人为标定类别什么的，无标注数据可以理解为原始的数据。有标注数据的话可以用来训练分类器，因为它相当于给出了函数的输入和输出值，来求一个函数或者概率模型，适合有监督学习，但是无标注数据可以理解为仅仅有输入数据，所以实际上只能分析这些数据的模式，从而进行进一步的研究。

1. 如何理解P248最后一段话？

自己的理解:

这里给出的是一个概率分布P(x|z)，z代表类别，而x代表样本，对于z来说，z有多个值，代表多个类别，而此时x固定，可以根据不同类别对应的概率大小将x分配到不同的聚类中，而不是像硬聚类那样是一个映射的关系。因为软聚类的定义里说一个样本可以属于多个类别，所以概率模型比较合适。

1. 读书计划

1、本周完成的内容章节：19章，20章前两节

2、下周计划：20章结束，看21章的前半部分

四、读书摘要及理解

1. 无监督学习基本原理

无监督需欸主要包括聚类，降维和概率估计，其对应着对数据进行纵向压缩，横向压缩以及对纵向横向结构的研究。

2. 研究问题:

聚类:将相似的样本点按照某种相似指标和制定的类别数进行聚类，使类似的点归于一个聚类，而不类似的点有明显区别。

降维:将低维的样本点通过降维方法压缩，称为高维空间的样本点，其要满足信息损失最小的原则，压缩了数据，保存了信息。

概率模型估计:假设获得的数据是由一个概率模型产生的，现在的目标是用数据去学习模型的结构和参数，因为模型的结构多种多样，一般要实现假设模型属于哪一种类别，这样比较好训练。而模型的具体参数通过学习给出。使用了贝叶斯的思想。

3. 机器学习三要素:

模型:概率还是非概率，决定了学习的结构。

策略:给出一个损失函数，进行最优化。

算法:如何高效进行最优化。

4. 无监督学习方法:

对应于三种问题，一般有着四种总的方法。

聚类:对样本点划定类别，一般其有着层次聚类和k均值聚类等具体的方法。

降维:降维的算法一般是使用PCA方法进行，而其要用到奇异值分解的基础知识。

话题分析: 概率模型估计的一种问题。其用来估计文本的话题，有着广泛的应用。

图分析: 概率模型估计的另一种应用，其用来分析图中的统计规律以及潜在结构。

5. 聚类:

聚类的问题首先要考虑相似度的定义，相似度的选择对于聚类的效果有着决定性的影响。一般有范数，相关性和夹角余弦。

之后，为了有效的定义聚类，我们要有一个指标来划定聚类的边界，决定哪些点可以属于一个类，其比较直观的就是定义一个距离，如果该聚类中任意一个点的距离都满足这个距离限制，则可以称之为一个聚类。

此外，为了分析不同聚类的关系，还需要定义聚类的中心的概念以及聚类之间的距离，常用的距离有最短距离，最长距离等。

思考:无监督学习方法与有监督学习方法有着明显的不同，有监督学习方法学习的是一种关系，而无监督学习方法学习的是原始数据中存在的模式和特征。其重要意义在于其可以在不标注的情况下对原始数据进行分析，找出其潜在的模式甚至是生成的模式，这对于未来训练新模型产生数据以及舍弃冗余数据进一步促进有监督学习有着重要的意义。