统计学习读书报告

09017244 郑健雄

1. 自己提出的问题的理解：
2. 常用的两种PCA求解方法之间有什么区别与联系？哪一种效率更高？

讨论后的理解：

两种算法都是求解协方差矩阵的特征值和特征向量，特征分解是直接求解，而svd则是构造一个新的矩阵来求解。SVD不需要计算矩阵相乘的步骤，保留了更多的信息，同时充分利用了协方差的性质，速度很快。

1. 主成分分析得出的结果除了可以用于聚类分析还有什么常见应用吗？

讨论后的理解：

一般可以用于需要数据降维的场合，比如说线性回归，SVM的分析都可以通过降维得到简化，再比如说NLP中词向量的求解也可以通过降维来获取。

1. 样本协方差矩阵S为什么是总体协方差矩阵的近似？

讨论后的理解：

该公式中除以1/n-1是因为这样才是总体的无偏估计。直接除以n的话虽然符合直觉，但因为假设总体均值和样本均值一定相等，很难满足，一般情况下是估计的结果偏小，不是总体的无偏估计。此处的推导用到了方差和期望的关系公式，得出的是数学上的无偏估计表示。

1. 别人提出的问题的理解
2. 书P313在用数据的协方差矩阵或相关矩阵求解PCA时，其实只要算到（4）就已经可以实现降维，后面那几步是不是可以没有，仅仅是为了对数据进行分析？第二种方法仅实现降维，后面几步并没有去算贡献率之类。

自己的理解:

后面的步骤是对降维质量的一个检验，也就是评估对于单个变量的信息比例。其即时对降维结果的评估，也是是否增大k值的依据，比如可以适度让k的值增大，从而保证单个变量的信息达到要求，而总体的信息量是不能一次保证分析的结果成立的。所以这几部分应该是必要的，如果不符合要求的话，可能需要重新规划参数和计算，奇异值分解那部分感觉像是省略了一些步骤，比如说k值的估计步骤实际上在15章中是svd计算会进行的

1. 书上介绍了两种求解PCA的方法，为什么现在常用第二种方法，是因为第二种方法效率更高吗？因为一个求解特征值和特征向量，另一个需要进行SVD 。

自己的理解：

可能主要是因为特征分解本身的局限性，特征分解只能用于方阵，并且必须满足对角化才能使用，而奇异值分解可以用在任何矩阵上，所以用途更广泛。

1. 第二种方法的k值必须提前指定吗？这个算法好像不像第一种可以进行选择。因为直接进行截断奇异值分解，事先就必须先将k确定。

自己的理解:

这个一种应该就是按照经验指定，然后直接计算验证。另一种思路应该是完全奇异值分解，然后按照保存信息的量或者弗罗贝尼乌斯范数的比例来直接筛选，导出前k列和前k行的信息，直接生成需要的矩阵。

1. 样本的协方差矩阵和相关矩阵分别代表着什么？

自己的理解：

都是对样本变量相关关系的度量，只是角度不太一样，不过如果数据进行过规范化的话，两者会统一为一个矩阵。

1. P315实例中给出了第一、第二主成分代表的含义，但似乎是事后从观察得到的，从实际意义上怎么理解主成分对问题的意义呢？

自己的理解：

主成分分析应该是按照重要性给出对于数据最重要的维度或者说特征，所以我们只知道哪些因素是重要的，算法并不知道它们的意义。而主成分的含义应该是我们观察其影响或者与数据之间的关系之后赋予的。如果首先不知道重要因素，也就说不上赋予意义了，所以PCA是必要的。

1. 读书计划

1、本周完成的内容章节：复习16章并阅读Speech and Language Processing第二章内容。

2、下周计划：准备复习17章并阅读第三章内容。

四、读书摘要及理解

1． 样本主成分分析：

总体主成分分析从概率论角度对该问题进行了建模，得到了比较完善的结果，但是其分析过程是建立在数据全面的理想情况下，现实中是很难实现的。为了处理这个问题，从数理统计角度进行了样本层面的PCA，其是基于样本估计总体的方法进行分析的，这种方法用到了随机生成数据，采样以及数理统计中常用的估计公式，对主成分分析中需要用到的参数进行了估计，其主成分分析过程和总体分析是一致的，并且也可以通过规范化处理来简化运算。

2. 计算方法：

主成分分析问题主要是要找出特征向量作为主成分，一种思路是特征分解，直接对协方差矩阵进行分解，这种方式直观但是有局限性，比如只可以作用于方阵并且必须可以对角化。更广泛的思路是SVD分解，其无需计算矩阵相乘，同时对于任何矩阵的协方差矩阵都适用，是更方便效率更高的算法。

思考：PCA是经典的降维方法，是基于SVD的一种延伸，在基于神经网络模型提出之前，该算法在很多降维的内容中有过使用。总体主成分分析是理想的情况，其给出了分析的思路和步骤。样本主成分分析是通过样本估计总体的方式近似使用了参数来分析，其分析步骤和总体一致。因为PCA的目的是找出特征向量，所以可以使用SVD和特征分解的方式进行，SVD的应用更加普遍。PCA可以用在很多领域，比如词向量的分析，SVM的数据降维，以及信息压缩等重要应用上。