方骏-2020年7月26日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：如何选取主成分的个数？

讨论后的理解：可以通过累计方差贡献率来解决，当累计方差贡献率达到规定的百分比以上，就可以认为这些主成分已经含有足够的信息了，就选取这些主成分就可以了。

1. 别人提出的问题的理解：
2. 读书计划

1、本周完成的内容章节：《统计机器学习》第十六章

2、下周计划：《统计机器学习》第十七章

四、读书摘要及理解

1、主成分分析是一种常用的无监督学习方法，这一方法利用正交变换把由线性相关变量表示的观测数据转换为少数几个由线性无关变量表示的数据，线性无关的变量称为主成分。主成分的个数通常小于原始变量的个数，所以主成分分析属于降维方法。

2、首先对给定数据进行规范化，使得数据每一变量的平均值为0 ，方差为1 。之后对数据进行正交变换，原来由线性相关变量表示的数据，通过正交变换变成由若干个线性无关的新变量表示的数据。新变量是可能的正交变换中变量的方差的和〈信息保存〉最大的，方差表示在新变量上信息的大小。将新变量依次称为第一主成分、第二主成分等。

3、假设为维随机变量，其均值为，协方差矩阵为。

考虑、由维随机变量到维随机变量的线性变换：



如果该线性变换满足以下条件，则称之为总体主成分：

（1）

（2）

（3）变量是的所有线性变换中方差最大的；是与不相关的的所有线性变换中方差最大的；一般地，是与都不相关的的所有线性变换中方差最大的；这时分别称为的第一主成分、第二主成分、…、第主成分。

4、假设为维随机变量，其协方差矩阵为，的特征值分别是≥≥…≥≥0，特征值对应的单位特征向量分别是, 则的第主成分可以写作：



5、样本主成分分析就是基于样本协方差矩阵的主成分分析。主成分分析方法主要有两种，可以通过相关矩阵的特征值分解或样本矩阵的奇异值分解进行。