**读书报告**

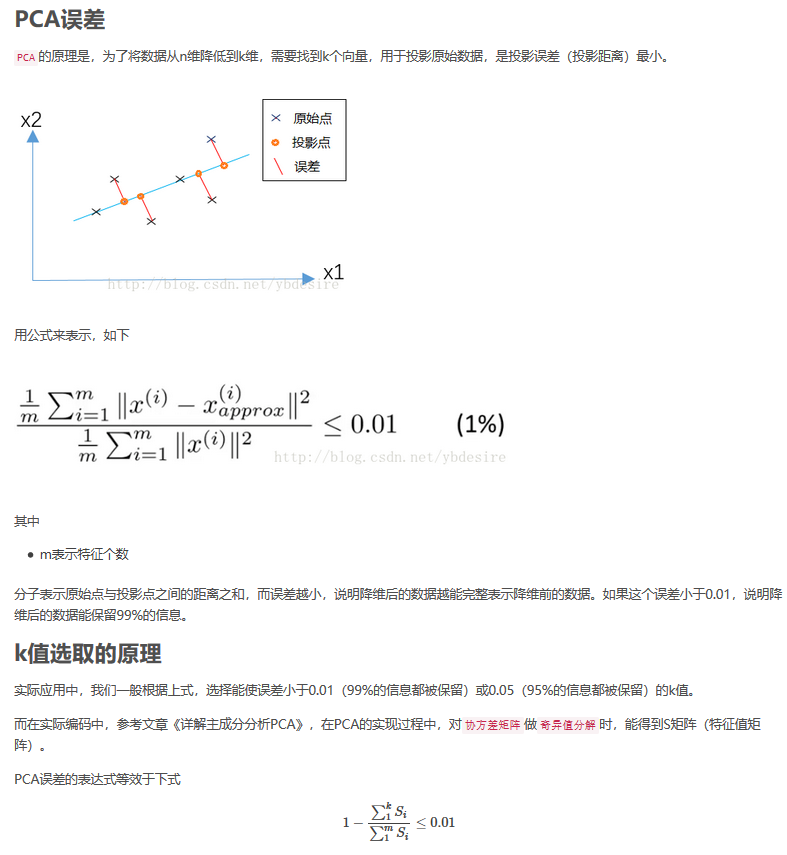
61518424 王贵涛

1. **问题与解答**

我提出问题：

1. 降维降的维数由什么决定？

讨论结果：这个可以人为进行选定，也可以划定一个标准来决定保存多少信息，比如70%的信息，主成分分析中可以计算出至少保留多少个维度的情况下可以保存这么多的原始信息，这种情况下就能计算出维数。



1. 非线性降维是如何实现的？

讨论结果：核化线性（KPCA）降维是一种非线性映射的方法，核主成分分析是对PCA的一种推广。KPCA主要利用了核函数，即对于当前非线性不可分数据，将其映射至更高维的空间至线性可分，再进行降维，而其中利用核函数可求得内积，进而得到样本在特征向量上的投影。

1. LDA是什么？

讨论结果：LDA是潜在狄利克雷分配，是第20章的内容，其可以用于文本分析。

别人提出的问题：

1. 聚类算法一般在什么情况下使用？具体有哪些使用实例。

我的解答：1 基于用户位置信息的商业选址；2 中文地址标准化处理；3 国家电网用户画像；4 非人恶意流量识别；5 求职信息完善；6 搜索引擎查询聚类以进行流量推荐；7 生物种群固有结构认知；8 保险投保者分组；9 网站关键词来源聚类整和；10 图像分割。

1. **下周计划安排**

看完第十四章并负责一部分的讲解。

1. **读书收获**

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（agglomerative）或自下而上（bottom-up）聚类、分裂（divisive）或自上而下（top-down）聚类两种方法。因为每个样本只属于一个类，所以层次聚类属于硬聚类

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并，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，之后将已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，对于给定的样本集合，开始将每个样本分到一个类，然后按照一定规则，例如类间距离最小，将最满足规则条件的两个类进行合并

如此反复进行，每次减少一个类，直到满足停止条件，如所有样本聚为一类。

聚合聚类需要预先确定下面三个要素：

1. 距离或相似度：

闵可夫斯基距离

马哈拉诺比斯距离

相关系数

夹角余弦

1. 合并规则：

类间距离最小

类间距离可以是最短距离、最长距离、中心距离、平均距离

1. 停止条件：

停止条件可以是类的个数达到闭值（极端情况类的个数是1)

类的直径超过阂值