****

数据科学基础

降维方法研究

小 组 成 员 陈鸿超（SY1806214）

李铎坤（SY1806219）

刘 颖（SY1806418）

编制时间： 2019年5月

目录

[1 SVD(奇异值分解) 2](#_Toc10555856)

[1.1 简介 2](#_Toc10555857)

[1.2 一般的SVD分解应用 2](#_Toc10555858)

[1.3 局限性 3](#_Toc10555859)

[2 随机投影 3](#_Toc10555860)

[2.1 随机投影简介 3](#_Toc10555861)

[2.2 Johnson–Lindenstrauss定理 3](#_Toc10555862)

[2.3 随机投影步骤 4](#_Toc10555863)

[3 实验设计 4](#_Toc10555864)

[3.1 整体实验设计 4](#_Toc10555865)

[3.2 结果 4](#_Toc10555866)

[3.3 SVD分析 4](#_Toc10555867)

# SVD降维

## SVD简介

SVD是一种矩阵分解，可以将一个矩阵分解成3个更小矩阵的乘积，其数学定义如下：

对于矩阵，r(A)=r，

A的正奇异值为：

奇异值分解：

存在酉阵和酉阵，使得

如图1.1所示：

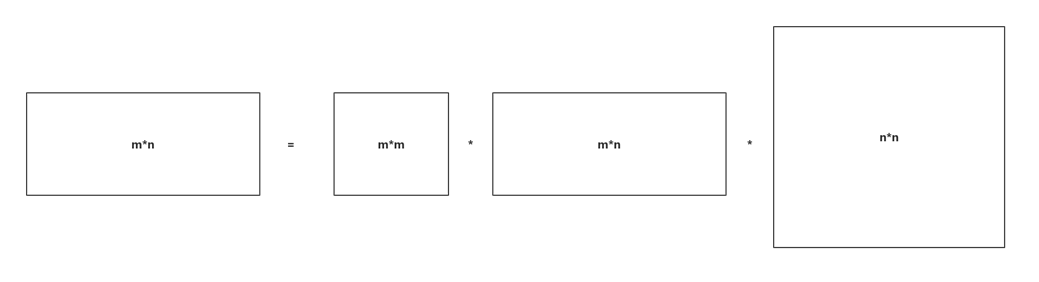


图 1.1奇异值分解

简奇异值分解：

存在半酉阵和半酉阵()，使得

如图1.2所示：

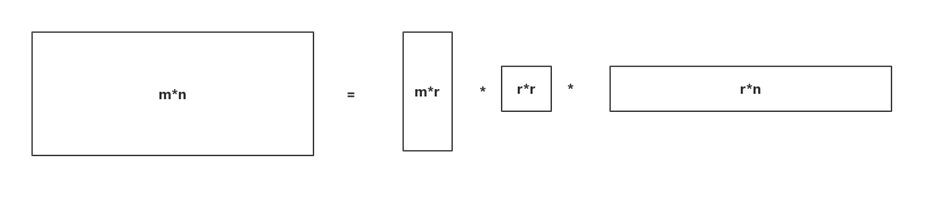


图 1.2简奇异值分解

## 传统的SVD降维方法

传统的SVD降维方法一般是将简奇异值分解的右矩阵作为降维后的数据，因为右矩阵的维度和原始矩阵数据中的数据个数相匹配。

比如在一个文本分类问题中，矩阵A是一个文本矩阵，每一列代表一个文本中各个词出现的频率，共n个文本，每个文本有m个词。其简奇异值分解如图1.3所示。

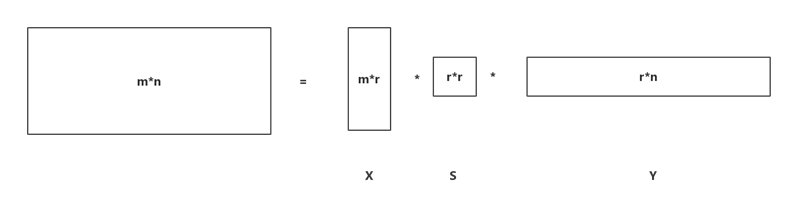
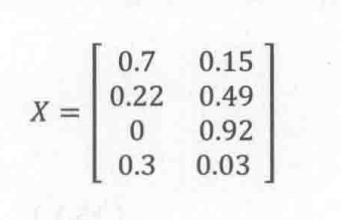
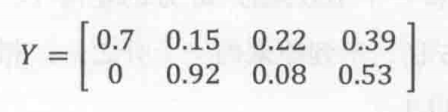


图 1.3文本分类问题的奇异值分解

其中，X是近义词分类结果，比如一个4\*2的矩阵X，代表着4个词，2个语义类，代表第i个词与第j个语义类的相关性。



Y是文本分类结果，比如一个2\*4的矩阵Y，代表2个主题，4个文本，代表第j个文本与第i个主题的相关性。



传统的SVD降维方法是一种黑盒操作，无法解释，同时也有很大的局限性，比如：

1. 降维后的维度不确定，也无法控制

即使是同一批数据，均分为两组，SVD分解后矩阵的维度都可能是不同的，相互之间的降维结果无法比较。

1. 降维效果不稳定

同一个问题，不同的数据、不同的特征、不同的特征表示，SVD降维的效果差别都可能很大。

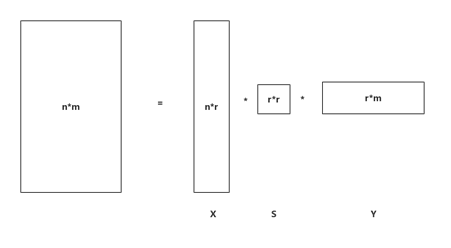
## 截断SVD降维方法

截断SVD是对传统SVD降维方法的改进，可以自由的指定降维后的维度，当然降维后的特征维度不能大于原矩阵的秩。除此之外，截断SVD降维方法还拥有很好的普适性，降维效果比较稳定，泛化能力较强。

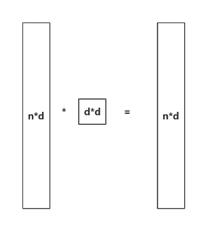
截断SVD降维方法认为，矩阵进行SVD分解之后得到的奇异值矩阵中，奇异值越大，代表着相关特征的重要性或者说有效性也就越大。因此，截断SVD的目的就是将原始矩阵中较大的若干奇异值与其对应的特征提取出来，组成降维后的数据。

截断SVD降维的步骤如下：

1. 对数据进行SVD分解，不同的是每一行是一组数据，共n组数据，每个数据有m个特征。



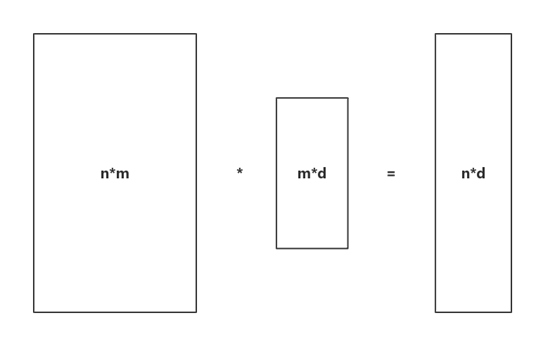
1. 因为奇异值矩阵S中的奇异值已经按照大小排序，所以直接取X的前d列与S的前d行d列相乘即可得到降维后的数据。



# 随机投影

## 随机投影简介

随机投影是一种数据降维方法，该方法会在原始数据的高维空间中随机选择一组单位向量，这组向量不必相互正交，然后将高维空间的数据映射到以这组单位向量为基的低维空间中。



随机投影如何保证降维效果呢？——Johnson–Lindenstrauss定理

## Johnson–Lindenstrauss定理

，，，

对

转换一下：

这条定理说明了以下几点：

1. 当投影维度确定之后，投影前后任意两点之间的距离差都是有上下限的，即投影效果是有上下限的。
2. 上下限取决于数据和投影维度，和投影方法无关。
3. 不同的投影方法只能让投影效果在这个界限内浮动。
4. 只要选择合适的维度，任意的投影方法都是有效的，可以接受的。

因此，只要确定合适的投影维度，随机投影和其他投影方法并不会有太大差别，投影效果都是可以接受的。同时随机投影最容易实现，是常用的降维方法。

## 随机投影步骤

随机投影流程一般可以分为4步：

1. 确认投影矩阵R的维度，即m和d的大小。
2. 使用随机数填充矩阵，常用的有高斯随机矩阵和稀疏随机矩阵。
3. 归一化矩阵
4. 对数据进行降维，Y=X\*R

# 实验设计

## 实验数据与实验设计

本实验准备使用上述提到的多种降维方法进行数据降维，以测试不同降维方法的效果。实验数据来源自模式识别课程，共3040张人脸图片， 152个不同个体，每个个体20张图片。

因为本实验的目的只是为了验证不同降维方法的效果，因此对于后续的分类操作只是简单的通过计算图片数据间的欧式距离决定图片的类别，并不准备使用复杂的神经网络，因此这会影响实验结果的分析，无法分辨分类效果的优劣是由降维方法导致的还是由于神经网络模型及其参数选择决定的。同时，本实验的评价指标也相应简化，只计算多分类结果的正确率。

本实验的操作流程如下：

1. 读取原始图片的数据
2. 将数据进行打乱
3. 使用降维方法对数据进行降维
4. 提取出每一类降维后的对照数据
5. 计算降维后的所有数据与每一类的对照数据之间的距离，对数据进行分类
6. 根据分类结果与原始标签计算正确率

本实验的流程图如图3.1所示。

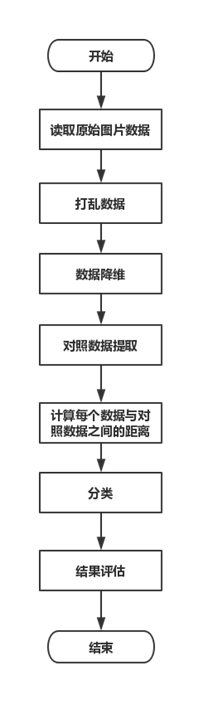


图 3.1实验流程图

实验数据的转换情况如图3.2所示。

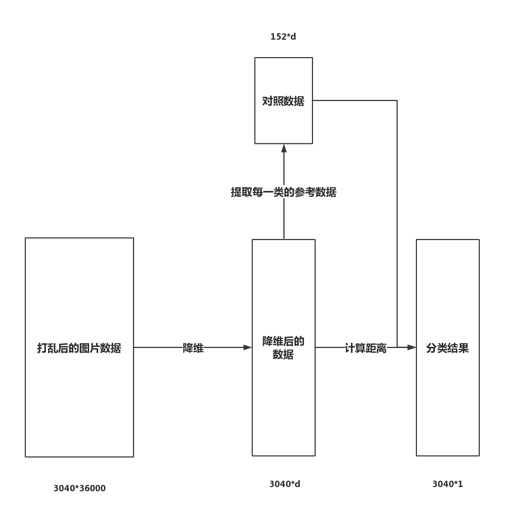


图 3.2数据转换图

## 实验结果

### SVD降维结果

本实验对传统的SVD降维方法和截断SVD降维方法都进行了测试，不幸的是传统的SVD降维方法对本实验的数据并没有产生效果，只有0.5%的正确率，即使加入了一些归一化和人为处理，效果也不太理想。

|  |  |
| --- | --- |
| **降维方法** | **正确率** |
| 不降维 | 92.87% |
| SVD | 0.5% |
| SVD+归一化+人为处理 | 最高10% |

相比之下，截断SVD方法的效果更加优异，并且经过多次实验，最优的正确率能够达到95%左右，最优的维度应该在16~64之间。

|  |  |
| --- | --- |
| **降维方法** | **正确率** |
| 不降维 | 92.87% |
| TSVD\_1024 | 92.76% |
| TSVD\_128 | 94.32% |
| TSVD\_64 | 94.88% |
| TSVD\_32 | 94.91% |
| TSVD\_16 | 93.28% |

### 随机投影降维结果

针对随机投影，本实验使用了高斯随机矩阵和稀疏随机矩阵两种构造随机矩阵的方法，而降维后的维度为6873维，该维度由Johnson–Lindenstrauss定理计算得到。

|  |  |
| --- | --- |
| **降维方法** | **正确率** |
| 不降维 | 92.87% |
| gauss | 93.07% |
| sparse | 93.04% |

### PCA降维结果

### 整体比较与分析

本实验采用的多种降维方法的效果如下表所示。

|  |  |
| --- | --- |
| **降维方法** | **正确率** |
| 不降维 | 92.87% |
| SVD | ≈10% |
| TSVD | ≈95% |
| 随机投影 | ≈93% |
| PCA |  |

除了SVD降维方法对本实验的数据无效以外，所有的数据降维方法在减少数据维度的同时，也起到了一定的特征提取的作用，提高了分类的效果。不过与神经网络相比，这种特征提取计算量较小，提取出的特征与用于计算的数据密切相关，并且没有固定的参数，无法学习。