**模式识别第四次作业**

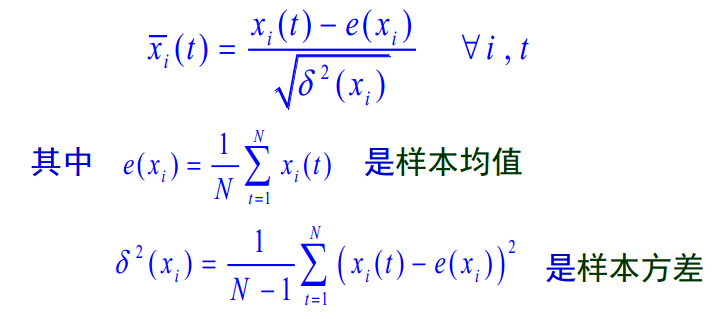
**用身高体重等数据进行聚类的实验**

2013011570 自31 唐静娴

1. **实验内容**
2. 用PCA对dataset3.txt样本的特征进行降维，画出各个主成分的方差，并据此确定若干主成分构成新的特征；
3. 在新特征表示的基础上，分别试验将训练样本聚成1、2、3、4、5、6类，绘制误差平方和随聚类数变化的折线。并据此确定最合适的聚类数进行聚类，用合适的方式表示出聚类结果并对其进行分析和讨论。
4. **实验方法**
5. **特征降维：主成分分析(PCA)方法**

主成分分析是数学上对数据降维的一种方法。其基本思想是设法将原来众多的具有一定相关性的指标重新组合成一组较少个数的互不相关的综合指标（主成分）来代替原指标。根据数学知识可以知道，每一个主成分所提取的信息量可以根据方差来度量，方差越大，包含的信息越多。通常情况下，数据中的大部分信息集中在较少的几个主成分上。本次作业指定**主成分所能代表的数据总方差比例为80%**，据此可以确定提取的主成分个数。其计算步骤如下：

1. 样本数据规范化。为了消除变量单位的影响，需要先将样本数据规范化，具体的做法为：



1. 计算协方差矩阵；
2. 计算协方差矩阵的特征矢量和特征值；
3. 选择成分组成模式矢量。从协方差矩阵中得到特征矢量后，便可以根据特征值的由大到小排列来得到成分的重要性级别。选择前m个成分，使得满足方差比例阈值要求，这p个成分便组成了我们需要的模式矢量pcs；
4. 获得新数据。

矩阵pcs为维，矩阵X为维。通过公式：

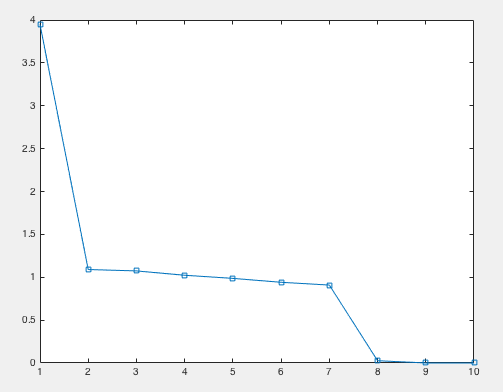
得到n个样本数据在m个主成分方向上的投影，即维的cprs\_data矩阵。

1. **聚类分析：C均值算法**

C均值算法也叫K均值算法，其目标是将样本划分为c类，使得总的误差平方和最小。计算步骤如下：

1. 选择C个点作为初始聚类中心；
2. 将每个点指派到最近的质心，形成C个聚类；
3. 重新计算每个聚类的质心；
4. 判断每个聚类的新质心与当前质心的偏差delta。若delta小于某个阈值，则认为聚类不发生变化，算法终止；否则以新质心替换当前质心，回到（2）。
5. **实验过程**
6. 指定方差的误差阈值为0.2，即主成分所能代表的数据总方差**不小于80%**，进行主成分分析，得到由5维新特征描述的样本数据cprs\_data；
7. 采用**随机选取（Random Partition）**的方法确定初始聚类中心，并用**欧氏距离（Euclid Distance）**度量样本点之间的距离，偏差delta阈值指定为0.001，对cprs\_data进行C均值聚类。分别取c=1,2,…,6,计算**误差平方和(WCSS, Within-Cluster Sum of Squares)**，并由此得出最佳聚类数目。
8. **结果分析**
9. **主成分分析**

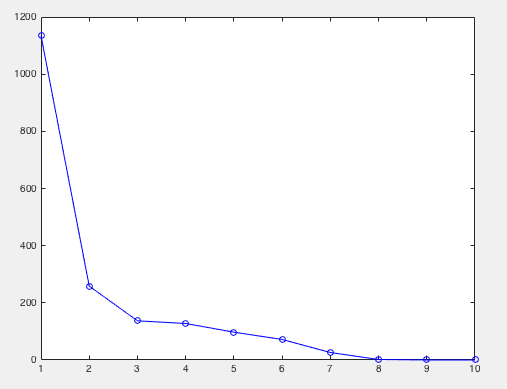
各个主成分上的方差为：



**图4-1-1 各主成分上的方差(规范化后)**

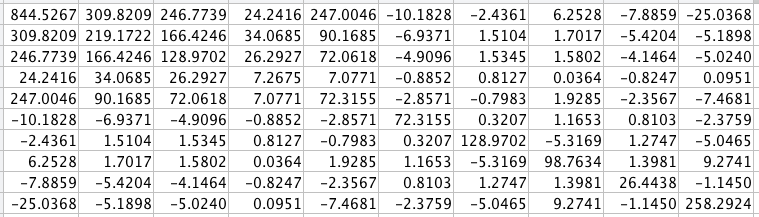
前5个主成分所含的数据信息已经超过了阈值80%，因此提取前5维主成分作为新的特征。

再来考虑不进行数据规范化预处理的主成分分析结果（图4-2），前3个主成分所代表的数据总方差达到89.36%。如果我们放松阈值要求至75%，甚至只需要2个主成分就可以解释原来的10维特征。利用因子载荷分析不难发现，此时对原始信息存在一定的丢失情况。



**图4-1-2 各主成分上的方差(未规范化)**

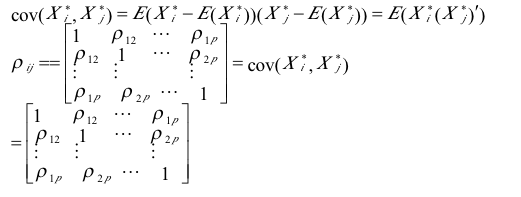
原始10维数据的协方差矩阵是：



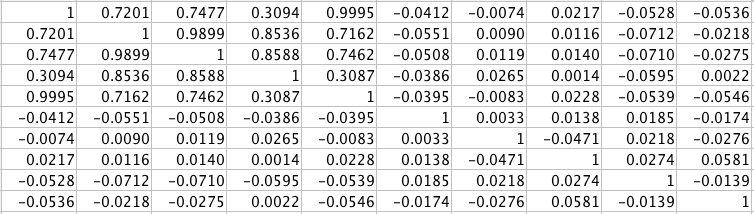
**图4-1-3 原始数据协方差矩阵**

数据规范化后，总体的协方差矩阵与总体的相关系数相等，证明过程如下：

（参考自网络）



因此通过计算规范化后数据的协方差矩阵，可以得到原始数据的相关矩阵：



**图4-1-4 原始数据相关矩阵**

协方差矩阵体现出前5维数据的离散程度较大，相关程度小；但相关矩阵却表明第1，2，3，5维数据之间有很强的相关性，相关系数达到了0.7以上。这与协方差矩阵给出的解释不尽相同，因为原始数据的各项指标受到了不同度量尺度的影响。本次作业中变量的量纲水平差异并不明显，因此还不能很好地体现出原始数据规范化处理的优点，但已经可以清晰地看到量纲对于主成分分析的影响。

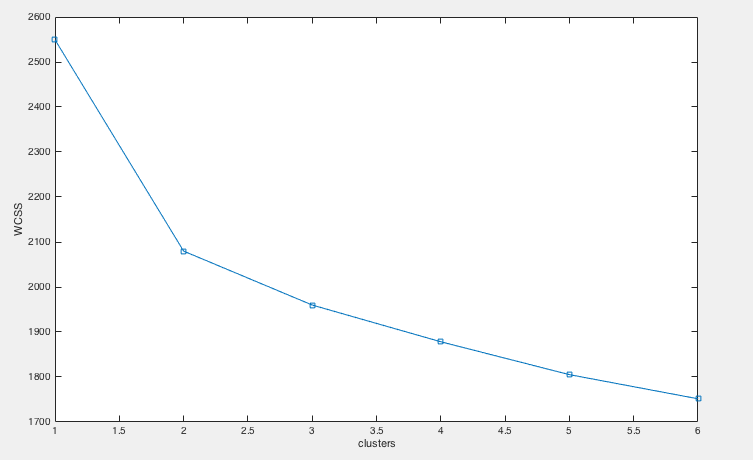
在实际应用中，我们应该在主成分分析之前先消除量纲的影响——对数据进行规范化处理，以使每一个变量的均值为0，方差为1.

1. **C均值聚类**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **C** | **1** | **2** | **3** | **4** | **5** | **6** |
| **WCSS** | 2549.42 | 2079.76 | 1959.91 | 1878.29 | 1805.25 | 1751.60 |

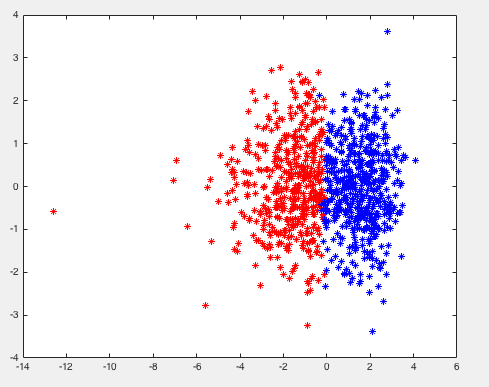
当C=1~6时，WCSS的值分别为：

折线图表示为：



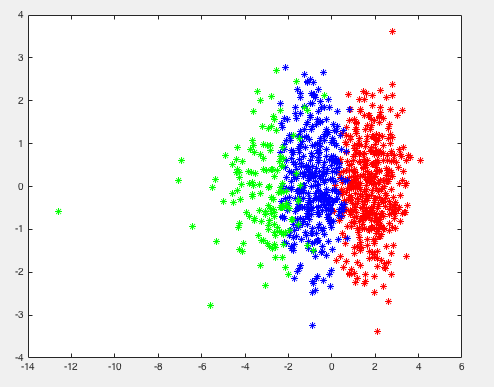
**图4-2-1 误差平方和曲线**

可以看到，当c=2时出现拐点，可以推断出最优的聚类数目为2.为了方便表示，这里只绘制出样本的前2维特征来表示聚类结果：

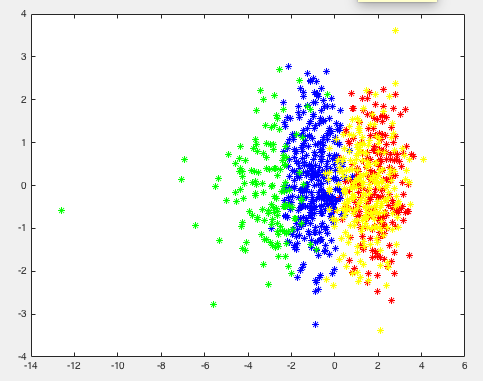
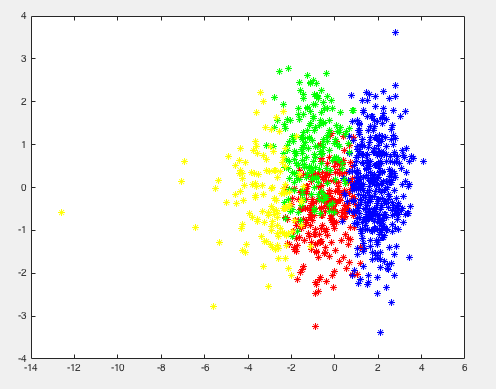


**图4-2-2 聚类结果（C=2）**

这与我们的直觉判断相同，最合理的聚类结果便是聚为“男”、“女”两类。再来看聚类数目更多时的结果（仅列举出一部分）：



**图4-2-3 聚类结果（C=3）**

**图4-2-4 聚类结果（C=4） 图4-2-5 聚类结果（C=4）**

随着聚类数目增加，前2维特征的重叠现象逐渐变得严重，虽然误差平方和在持续减小，但已没有实际意义。注意到当聚类数目C=4时，出现了两种聚类结果，印证了“C均值算法收敛到局部最优”这一说法。

1. **程序来源说明**

本次作业代码全部自行编写，开发平台为Matlab R2015b.