模式识别第四次作业

实验目的：

学习 PCA 降维方法和聚类方法，并进行实验，并对聚类效果进行讨论。并在过程中学习非监督学习的方法。观察他们的效果。

实验方法：

一、对原数据进行 PCA 降维。
二、利用 K-MENS 对特征进行分类。

实验原理：

1. PCA

对原始特征进行线性组合

$$\xi_i = \Sigma_{j=1}^{p} a_{ij} x_j$$

希望该重新组合的新的特征方差最大

$$\mathrm{var}(\xi_1) = \boldsymbol{a_1^T \Sigma a_1}$$

加上约束条件

$$\mathbf{a_1^T a_1} = 1$$

得到拉格朗日函数，求其极值

$$f(\mathbf{a_1}) = \boldsymbol{a_1^T \Sigma a_1} - v\,(\boldsymbol{a_1^T a_1} - 1)$$

$$\boldsymbol{\Sigma a_1} = v\boldsymbol{a_1}$$

得到

$$\mathrm{var}(\xi_1) = v$$

于是我们取方差较大的前 N 个重新组合的特征，在该特征意义下，对原来的数据进行变换。

2. C 均值

$$m_i = \frac{1}{N}\Sigma_{y\in\tau} y$$

$$J_e = \Sigma_{i=1}^{c} \left|\left|\boldsymbol{y} - \boldsymbol{m_i}\right|\right|^2$$

目的是通过迭代，得到J_e的最小值。具体算法书中有。

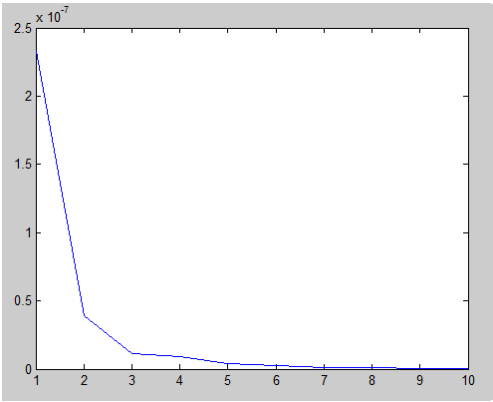
实验结果：

Task 1: PCA 方法构成新的数据表示

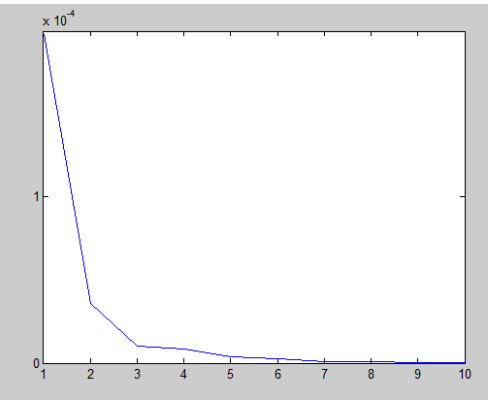
把数据dataset3.txt 作为未知样本集（即忽略每个样本的类别信息），用PCA 对dataset3.txt 进行降维，画出各个主成分上的方差，根据方差分布确定选取几个主成分来构成数据新的表示。

实验数据：

（横轴为 PCA 方法形成的新的 10 个特征，纵轴为该特征的方差）



图一：二范数归一化



图二：一范数归一化

讨论：

1. **对于原始数据需要进行归一化。**

   观察原来的数据，会发现其中有的组的数据，例如身高，本身较大，根据公式可以看到，这会产生较大的方差，因此，我们需要进行归一化。

2. **不同归一化条件下，对源数据进行 PCA 降维后得到的方差的值有些差距。**

   但是整体趋势相同，这与我们的理解相同，观察 PCA 方法推导，我们关注的是线性变换后的方差，虽然经过了不同的归一化，但是数据本身的方差的特征依然较好地保留在了归一化后的数据中。

3. **得到了方差，我们需要确定需要选择几个特征，有不同的标准。**

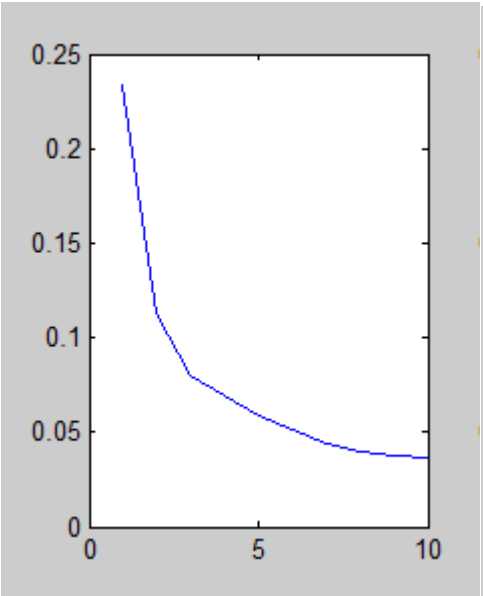
   在选择一范数归一化的情况下，分别计算选择 3,4,5 个特征所占比例为 93.61%，96.86%，98.20%。如果我们选择的标准定为 98%的话，我们选择保留 5 个特征。但是考虑到 5 个特征没有办法进行绘图直观表示，因此也选择三个特征进行计算。

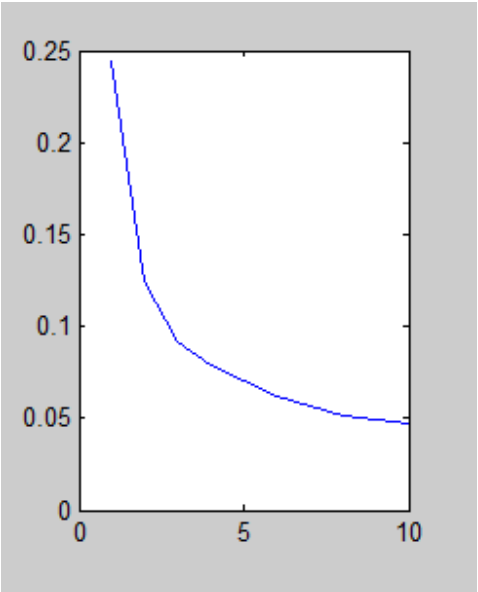
Task 2

在 task1 中，我们决定，在二范数归一情况下，分别保留三个特征和五个特征进行讨论。然后讨论 C = 2，3，4，5，6 的情况。

其中重点五个特征和三个特征时，C = 2 的状态，并与实际的情况进行比较，观察无监督分类的分类结果，并与前面几次作业的有监督分类得到的结果进行比较，并评价无监督分类的结果。

一、保留三个特征和五个特征，不同 C 的分类的方差变化。





图一：保留三个特征 **图二：保留五个特征**

（要求画 2——6，我为了观察整体情况画了 1——10）

根据这两张图我们可以看出

1.1、当保留三个特征的时候，方差较小。

这也符合我们的预期，PCA 本质上也可以作为数据压缩的一种手段。当选取的特征较少的时候，就意味着我们忽视了每组数据的末尾波动部分（有时称之为“噪声”），这就代表着误差平方和的下降。在图中我们可以明显地看到这一点。

1.2、随着聚类的增多，误差平方和下降

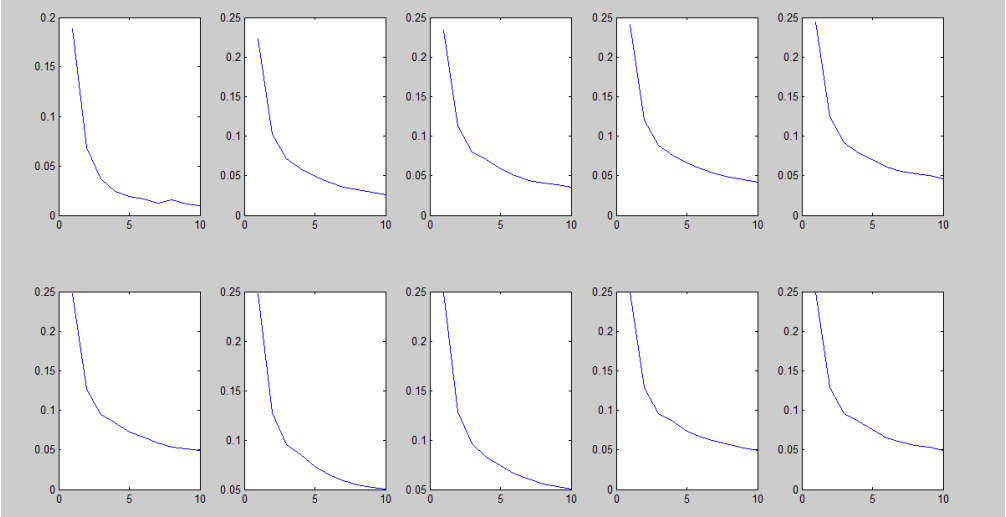
这也是较为直观的，因为有更多地类别代表着更好地适应性。但不一定适合我们的任务，因此需要判断聚类的数目。

二、保留 1~10 的特征，观察不同的 K-means K 值的不同，随之的变化曲线

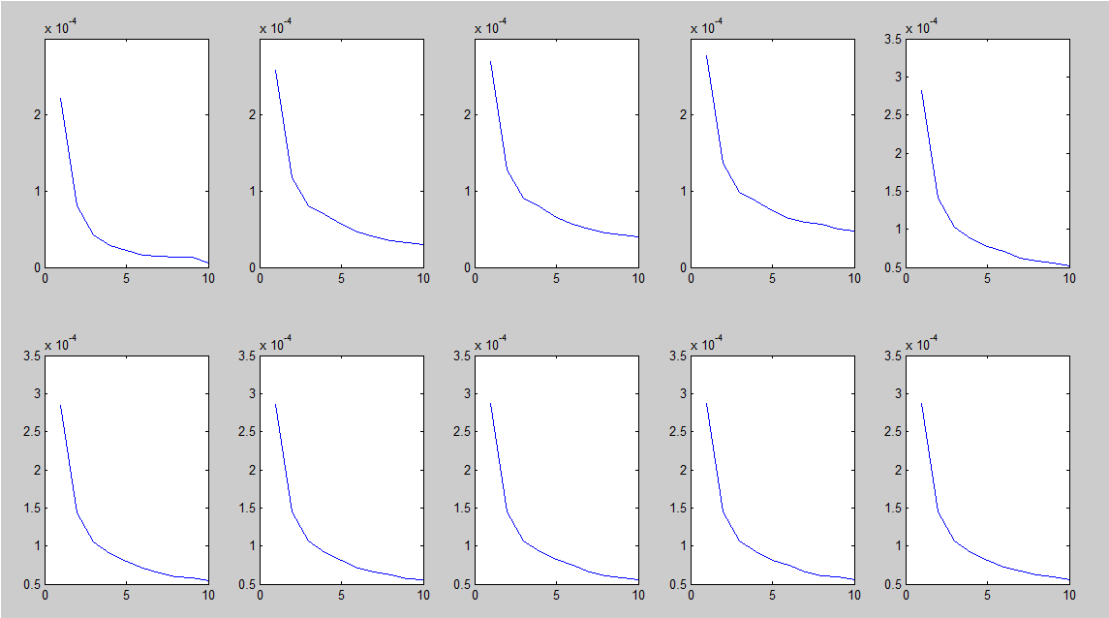
画出这样的图，仅是为观察取不同数目主特征带来的影响。

整体的趋势符合“一”中的关系。可以看到，当特征取 5 个以上时，原图已经基本不发生改变（注：第七和第八个图形状不同主要是因为坐标轴的选取不同）。

这也验证了我们先前“选前 5 个主成分”特征可以包含 98%以上信息量的论断。



图一：（二阶）正交归一后的结果



图二： **（一阶正交归一后的结果）**与图一结果相似，不做进一步探讨。

三、考察 PCA 分出的类别的特征

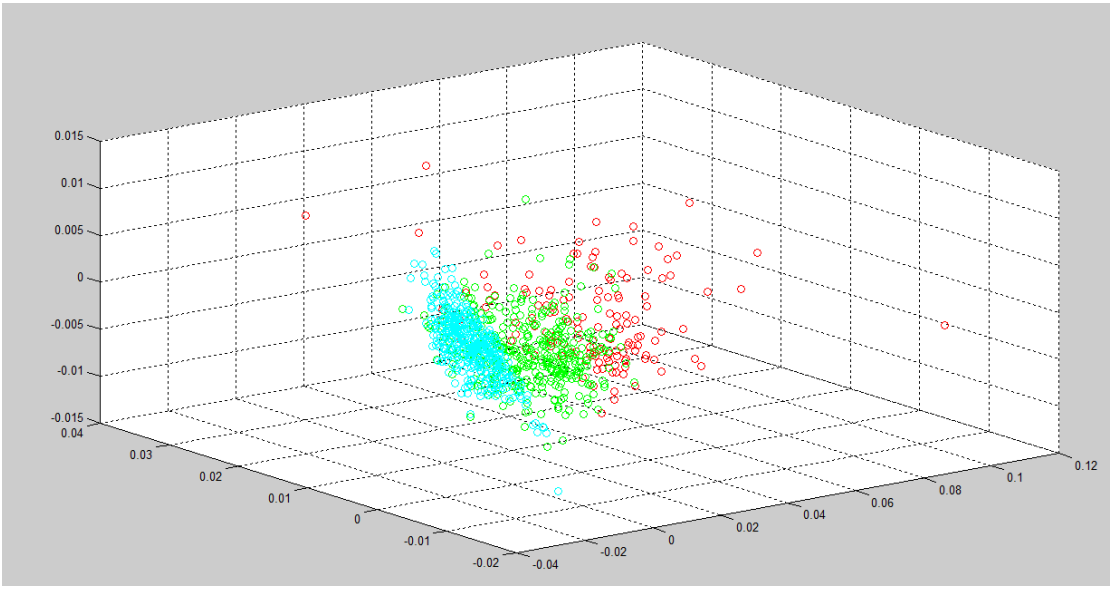
可以看到，保留三个特征和五个特征时，都在分类为 2 和 3 处出现了一个较为明显地拐点。之所以要使用这个拐点。是因为数据的聚类特征，设聚类的中心点为 C_1, C_2，C_3，如果有较好地聚类特性，此时若新增一个点则该点必然在 C_1，C_2，C_3 周围附近，否则平方差之和会大幅度下降。这就是为什么我最后会选择$C^* = 3$，进行讨论。

3.1 三个特征

这里，我首先观察了原始数据的划分状况，**但事实上，他们都可以对样本进行较好地聚类，因此并不能直接通过观察计算他们之间的距离。因此，在后面，我都结合真实情况对结果进行了检验，以评价这种方法的优越性。**

3.1.1 分三类

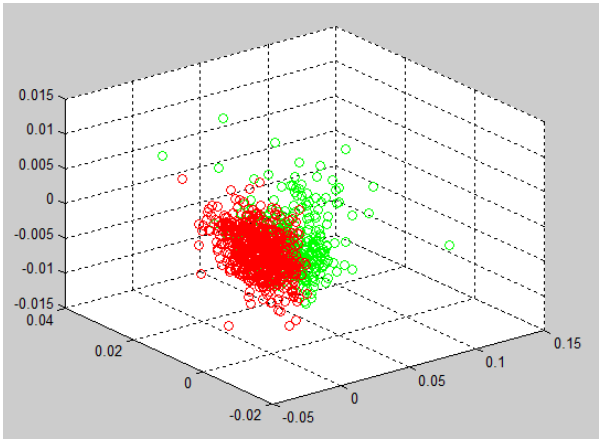
此时仅仅有三个特征，因此我们可以通过散点图的方式将分类的效果展现出来。得到的结果如下图所示：



5 feature	man	women
1	11	130
2	396	41
3	76	298

3.1.2 分成 2 类，

可以看到，出现了较为明显的聚类特征。



5 feature	man	women
1	451	162
2	34	307

观察图片可以看到，两类样本出现了较为明显的聚类特征。训练集错误率，20.5%,基本与之前的 fisher 方法相似。这个不难理解，因为 fisher 分类法寻找的决策面正是要求类间距离最大，类内距离最小，而 PCA 本质上也是尊崇着同样的道理。他们都属于非监督学习方法，但显然，比起监督学习（例如 MLP），非监督学习的结果要弱于监督学习。

3.2 五个特征，

五个特征没有办法画图。于是通过表格形式来表达结果。

3.2.1 分三类

5 feature	man	women
1	417	48
2	8	112
3	60	309

与三个特征相比较，男性的却分度上升，女性的区分度下降。但我认为本质上并没有提高，改变的仅仅是边界的点。

3.2.1 分两类

5 feature	man	women
1	34	307
2	451	162

发现，它与只选取三个特征时分类的结果是相同的。也是由于特征越靠前，包含信息量越大的原因。

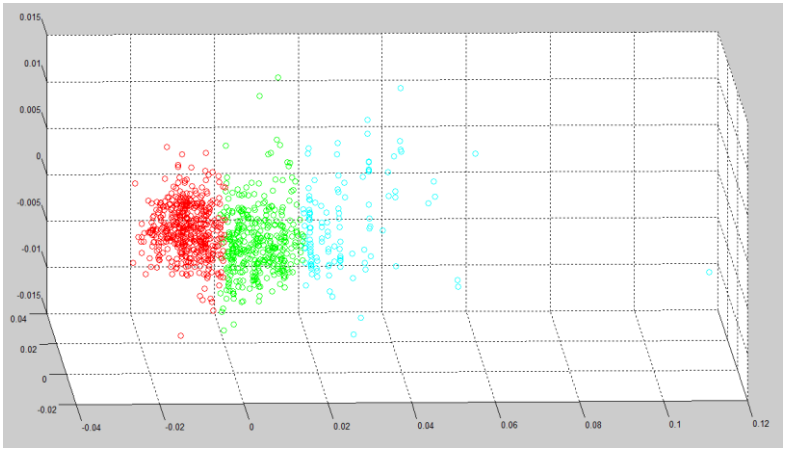
进一步的讨论：

我们会发现，无论我们将其分为两类还是三类，呈现出了聚类的效果。除了上面有关“不同判别法比较”，“不同特征数目带来结果的影响”等方面的讨论。再讨论一下究竟应该分成几类。可以看到，无论时两类还是三类，男性都呈现除了较好的聚类效果，而女性的聚类效果较差，实验进行到这里是无法得到这个问题的答案的，因为我们不知道原始数据的实际的分布情况。于是，对于 2b 部分的讨论，可以让我们从原始数据的角度出发去分析实验结果。

四、对于 PCA 变换后的 feature 聚类和原特征聚类的比较

4.1 三分类

4.1.1 绘图表示



从某个角度看过去，三者被较好地区分开来。但我们可以发现，主要的元素主要集中在，红色和绿色部分，蓝色部分较少，且较为分散，因此，我们可以考虑，分成两类可能是一种比较好的选择。

4.1.2 与原始数据分类情况的比较

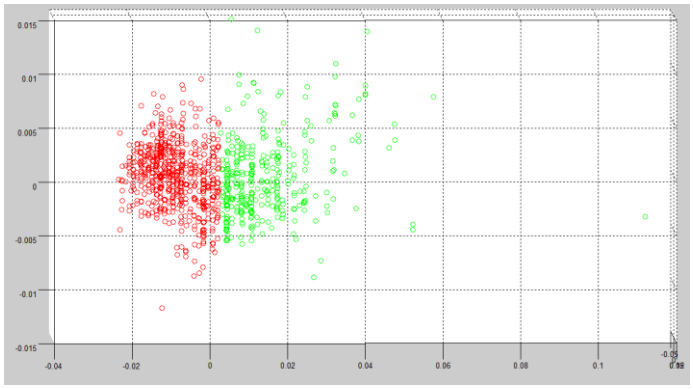
3 feature	man	Women	3(raw) feature	man	women
1	11	130	1	18	131
2	396	41	2	389	45
3	76	298	3	78	293

该表格左边为利用 PCA 选取的 3 个 feature（主成分），对样本进行的区分。
该表格右侧为利用 Raw DATA 其中较好的 3 个 feature（第三次作业得到的），对样本进行区分。

可以看到这两者几乎是一致的，因此无需再做图比较。这个也很好解释，PCA 的目的是选取信息量较大的特征：
1. 这三个特征由于自身可分性较强，因此这两者是相似的也不难理解。
2. raw Data 中剩余的特征可以在 PCA 的主特征中理解为“noise”

4.2 三分类

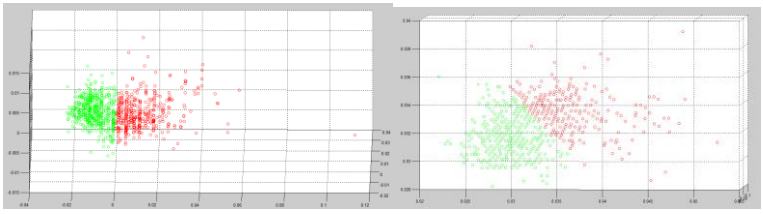
二分类（与三分类类似，不再讨论）



这印证了我们在三分类中得到的结论，两类是一种较好的分类方法

PCA	man	women	Raw	man	women
1	451	162	1	48	328
2	34	307	2	437	141

五、Abs 距离（即 cityblock 距离）下，PCA 与 RAW material 的差别



PCA	man	women	**RAW**	man	women
1	405	58	1	48	328
2	64	427	2	437	141

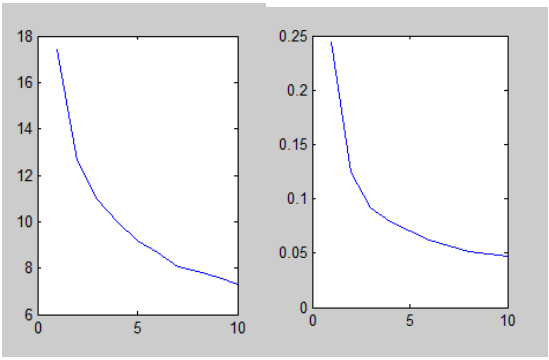
可以看到，当为 cityblock 时候，PCA 方法更好的分割效果。

对于 raw material 我们需要先判断该特征的可分性，对于 PCA,直接对源数据进行处理。这里我们看到了 PCA 方法的优越性。

六. 对于 PCA， abs 的距离和欧拉距离，cosine 距离的比较

6.1 三个特征

左侧是 abs 距离，右侧是 euler 距离，两者的距离取法不同，因此比较纵轴（distance）的绝对值的大小没有比较的价值。实际上，在这个问题中，这两种方法差别不大。



接下来，我们根据结果来衡量这两种方法的区别。

abs	man	women	Euler	man	women
1	405	58	1	451	162
2	64	427	2	34	307

3 feature	man	Women	3(raw) feature	man	women
1	21	222	1	11	130
2	389	31	2	396	41
3	75	216	3	76	298

可以看到，两种方法的差异不大。

6.1 五个特征

abs	man	women	Euler	man	women	cosine	man	women
1	432	110	1	451	162	1	451	105
2	53	359	2	34	307	2	54	364

abs	man	women	Euler	man	Women	cosine	man	Women
1	71	256	1	60	309	1	51	85
2	397	36	2	8	112	2	38	306
3	17	177	3	417	48	3	396	78

根据上面的四组结果，我们看出，在这个问题里，用哪种距离的影响并不大,且初值的选取可能也会对结果造成一定的影响。

在实际问题中，我们才需要考虑这个问题。

总结

在这次实验中，我对 PCA 方法有了更进一步地更深层次的理解。特别是在学习完特征的可分性判据后。我感到 PCA 方法是一种很好的具有很强生命力的非监督分类方法。在 C 均值部分，我探索了，不同的主成分、不同的距离、不同的类别、原始数据与 PCA 数据，对于最后结果的影响，但是由于直观方法不容易判断 ，我结合对样本信息的了解，对 PCA 方法进行了评估。

不同的标准，会产生不同的评估，因此，在实际运用中，我们需要根据实际情况做出正确地选择。

另外还发现，以上内容（距离、选择主成分个数等）虽然会对结果产生一定的影响，但是除了“原始数据与主成分数据”和极端情况有时会产生较大差异外，其他对于 PCA 的影响不大，因此，在未来的使用 PCA 过程中，我们在选择适合实验的参数后，不应该过分拘泥于个别参数的选取。

参考:《模式识别》（第三版）张学工 清华大学出版社

代码来源：PCA 自己写的，Kmeans 利用了 matlab 工具包。