

# 计算机科学与技术学院可视化技术课程实验报告

实验题目：可视化数据降维		学号：201900180065
日期：2021. 10. 15	班级：19 智能	姓名：倪诗宇
Email：1026507192@qq.com		

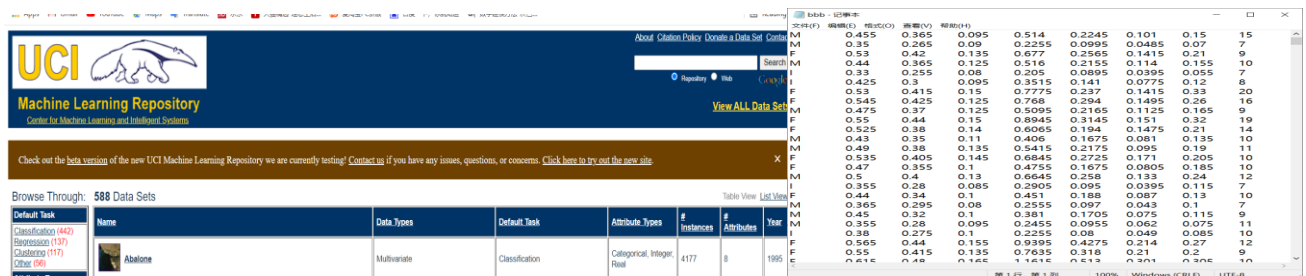
实验目的：  
体会不同降维方法对数据的降维效果

实验软件环境：

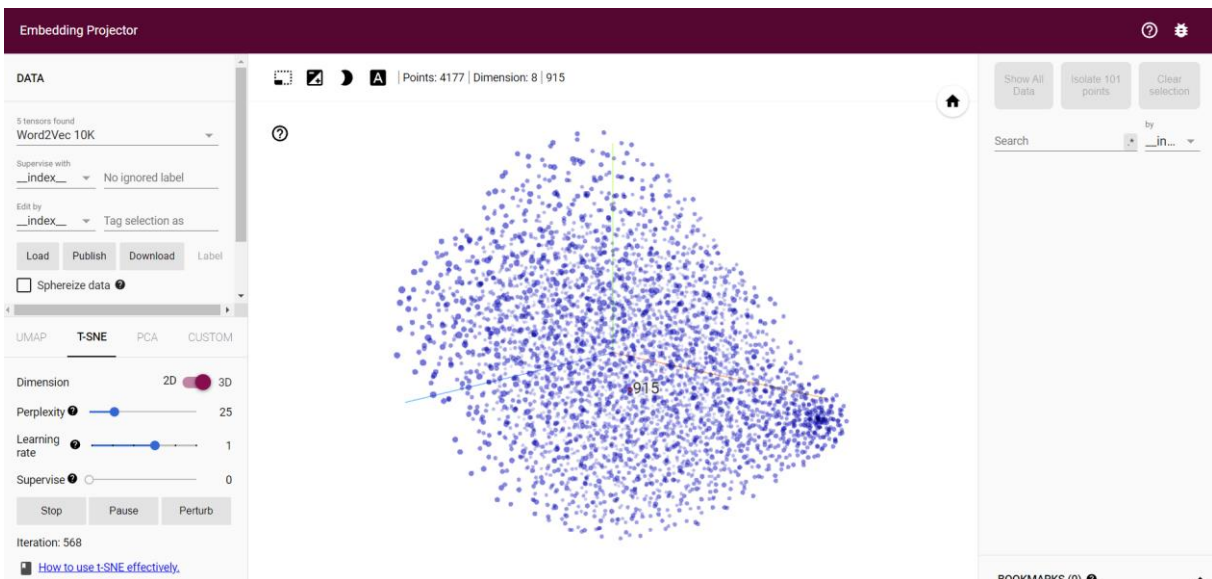
1. Matlab 2020 ， Embedding Project

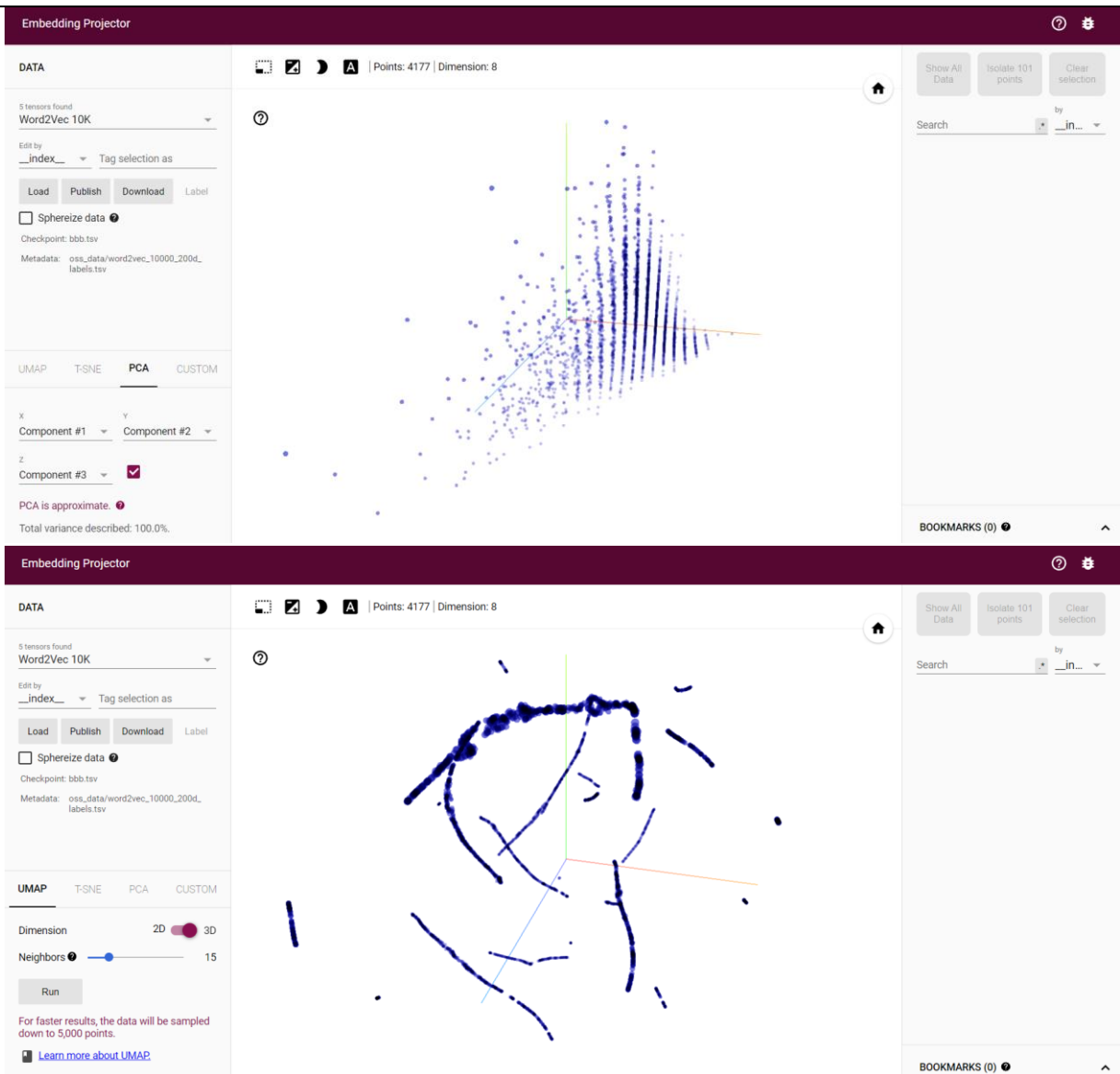
## 实验一

### 1. 下载数据 Abalone Data Set



### 2. 转成 tsv 文件格式，并上传





## 实验二

### 1.使用实验一中使用过的 Abalone Data Set

文件(F)	编辑(E)	格式(O)	查看(V)	帮助(H)				
M	0.455	0.365	0.095	0.514	0.2245	0.101	0.15	15
M	0.35	0.265	0.09	0.2255	0.0995	0.0485	0.07	7
F	0.53	0.42	0.135	0.677	0.2565	0.1415	0.21	9
M	0.44	0.365	0.125	0.516	0.2155	0.114	0.155	10
I	0.33	0.255	0.08	0.205	0.0895	0.0395	0.055	7
I	0.425	0.3	0.095	0.3515	0.141	0.0775	0.12	8
F	0.53	0.415	0.15	0.7775	0.237	0.1415	0.33	20
F	0.545	0.425	0.125	0.768	0.294	0.1495	0.26	16
M	0.475	0.37	0.125	0.5095	0.2165	0.1125	0.165	9
F	0.55	0.44	0.15	0.8945	0.3145	0.151	0.32	19
F	0.525	0.38	0.14	0.6065	0.194	0.1475	0.21	14
M	0.43	0.35	0.11	0.406	0.1675	0.081	0.135	10
M	0.49	0.38	0.135	0.5415	0.2175	0.095	0.19	11
F	0.535	0.405	0.145	0.6845	0.2725	0.171	0.205	10
F	0.47	0.355	0.1	0.4755	0.1675	0.0805	0.185	10
M	0.5	0.4	0.13	0.6645	0.258	0.133	0.24	12
I	0.355	0.28	0.085	0.2905	0.095	0.0395	0.115	7
F	0.44	0.34	0.1	0.451	0.188	0.087	0.13	10
M	0.365	0.295	0.08	0.2555	0.097	0.043	0.1	7
M	0.45	0.32	0.1	0.381	0.1705	0.075	0.115	9
M	0.355	0.28	0.095	0.2455	0.0955	0.062	0.075	11
I	0.38	0.275	0.1	0.2255	0.08	0.049	0.085	10
F	0.565	0.44	0.155	0.9395	0.4275	0.214	0.27	12
F	0.55	0.415	0.135	0.7635	0.318	0.21	0.2	9
F	0.615	0.48	0.165	1.1615	0.513	0.301	0.305	10

本数据中，第一列代表 Male，Female，Infant，为特征列。最后一列为年轮的圈数，为 label 列。因此，我们在进行降维的时候，将第一列的字符串分别转换为数字 0 1 2，将最后一列去除。然后用所得到的数据进行降维。

### 降维数据维度 4177 \* 8

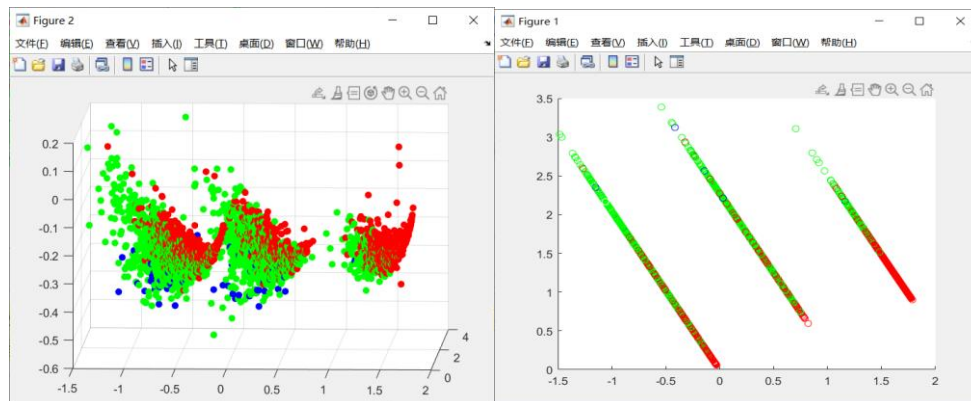
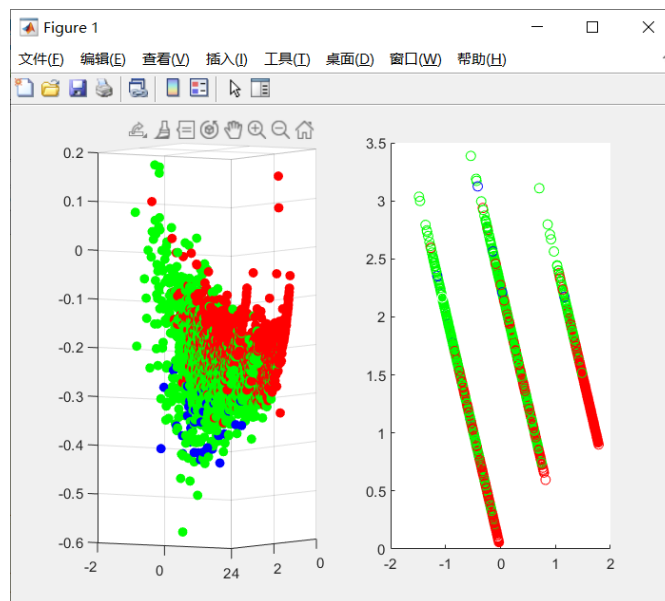
因为 label 的值都是整数，我们将一定范围内的 label 设置为同一种颜色。

0-10 为红色，10-20 为绿色，其他的为蓝色

1.1 利用 PCA 进行降维，分别降维到三维和二维，得到结果如下。

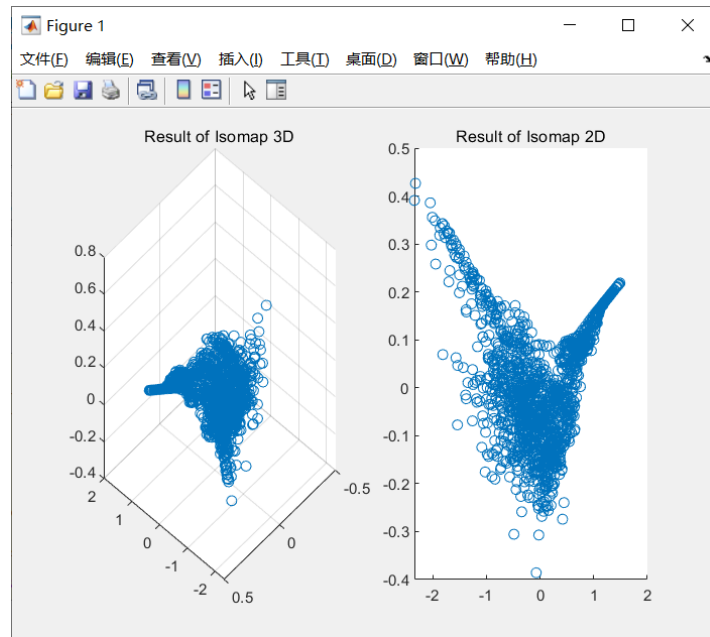
可以清晰地看到，红色和绿色分据两半，蓝色数据较少。

可以得知，鲍鱼的年龄（也就是年轮）主要集中在 0-20 岁之间，数据主要分成三个类别。且 0-10 岁鲍鱼的特征与 10-20 岁鲍鱼的特征有明显的差异，但也有部分重叠

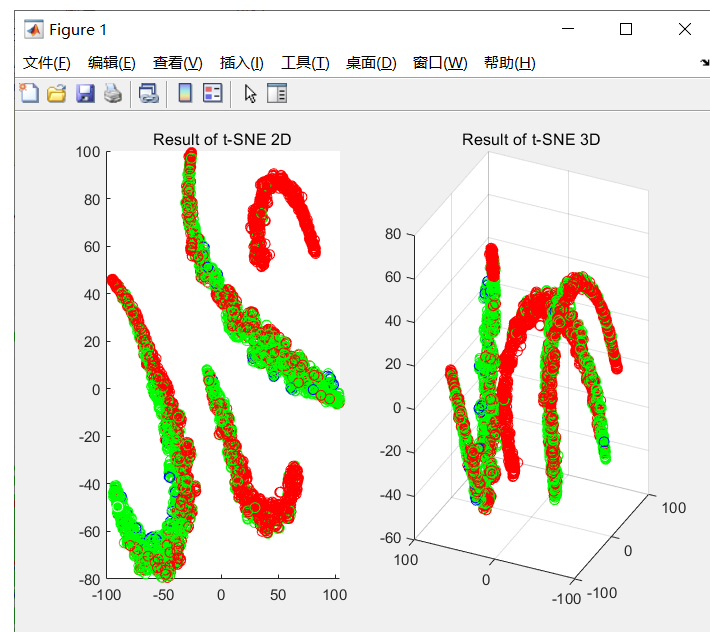


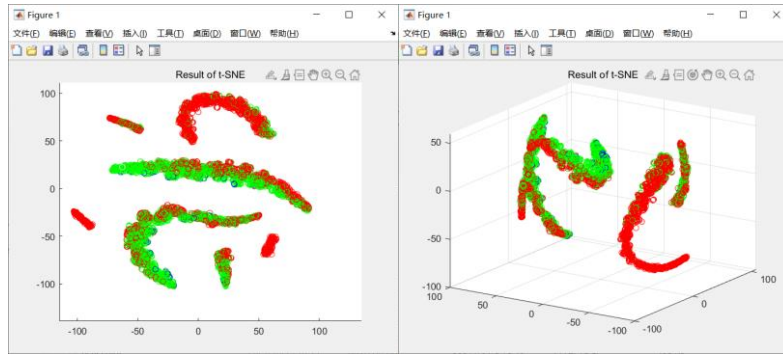
1.2. 然后我们进行 Isomap 降维，这里的降维出现了问题，降维后样本数变少了，至今还不明白原理。还望解答，很多同学都遇到了这种现象，从 4177 \* 8 降维至 1528 \* 2 / 3

attr	4177x8 double	4177x8
data	1x1 struct	1x1
i	4178	1x1
mappedX	1528x2 double	1528x2
mapping	1x1 struct	1x1
sex	4177x1 double	4177x1



这里由于数据量出现了减少，因此无法对不同类别进行颜色标注。无法获得有效信息  
1.3.最后是使用 t-SNE 进行降维，t-SNE 需要迭代的次数较多，运行比较慢



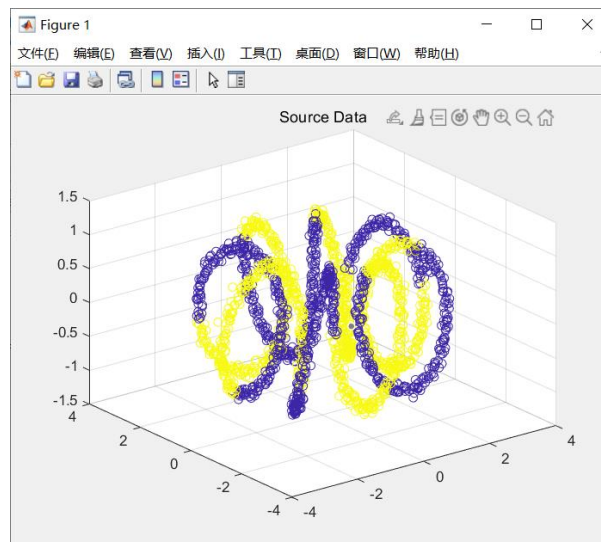


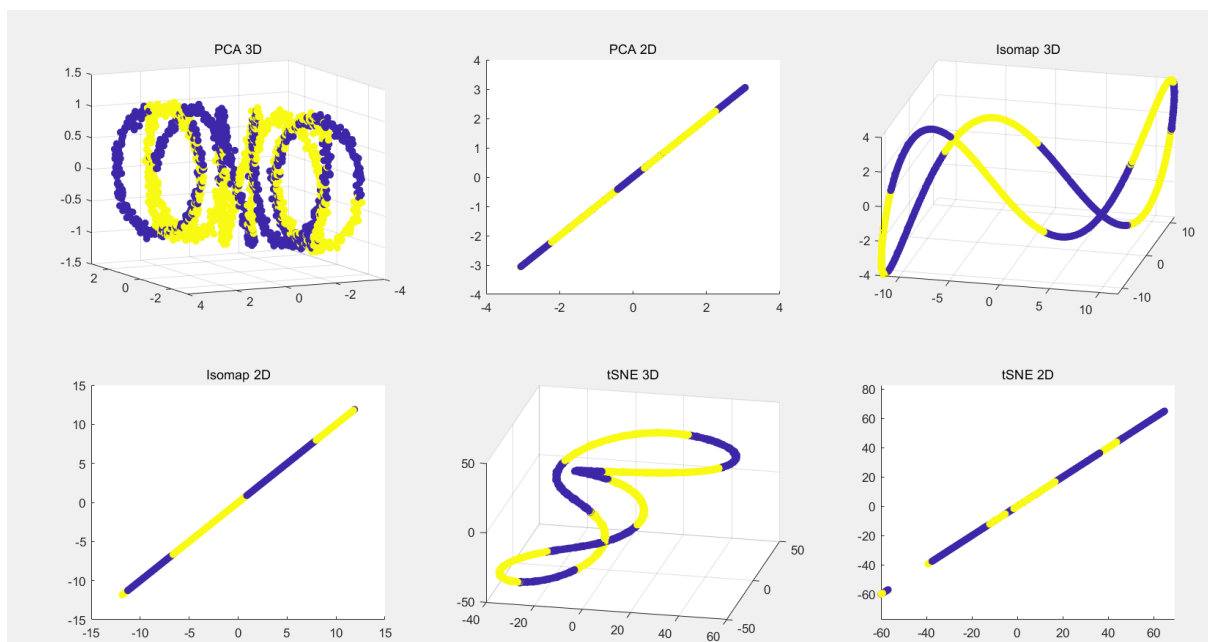
### 3. 使用 drtoolbox 的样例数据

左边的为特征  $2000 \times 3$ ，右边的为 label 为  $2000 \times 1$ ，且只有 0 和 1

这个数据集的 Isomap 并不会减少样本集

2.9343	0.0461	0.0368								0							
2.9770	0.0249	0.1553								0							
3.0142	0.0851	0.0315								0							
3.1736	0.0033	0.1978								0							
3.1302	0.0706	0.1021								0							
2.9206	0.0707	0.1176								0							
3.1356	0.1352	0.2055								0							
3.0152	0.0076	0.2352								0							
2.9702	0.0841	0.2383								0							
3.0029	0.0959	0.1753								0							
2.9500	-0.0148	0.2384								1							
2.9466	0.1738	0.2536								1							
3.0191	0.2608	0.2953								1							
3.0063	0.1176	0.3055								1							
2.9974	0.1524	0.3894								1							
2.9501	0.1235	0.4420								1							
2.8456	0.1746	0.4018								1							
2.9306	0.1449	0.4648								1							
2.9645	0.1672	0.4975								1							

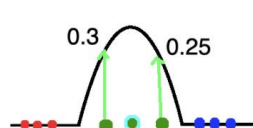




由上图可以看到，我们使用三种方法分别进行降维，显示二维，三维数据。二维数据都比较容易看，数据明显分成两份。在三维数据中，除了 PCA 以外，都很清晰地能看到数据被有界分为两类。

#### 结论分析与体会：

1. PCA 能有效处理线性分析问题，比如可以线性分开的数据，PCA 的降维效果比较好
2. PCA 在流型体上的降维效果不好。Isomap 可以对流型分布的数据进行降维，且不用计算测地线距离。使用的方法为构建连通图，每个点只和距离其最近的 K 个点相连，构建邻接矩阵，进而求出任意两点之间的距离，来代替 MDS 中的欧氏距离
3. t-SNE 将距离表示为概率，相似的样本由附近的点建模，不相似的样本由高概率的远点建模。t-SNE 是当前最好的降维方法之一



$$\frac{0.3}{0.3 + 0.25} = 0.545$$

$$\frac{0.25}{0.3 + 0.25} = 0.454$$



$$\frac{0.125}{0.15 + 0.125} = 0.454$$

$$\frac{0.15}{0.15 + 0.125} = 0.545$$

4. Isomap 降维时可能会使某些数据丢失。
5. 降维，尤其到低维，有利于数据分布的观测