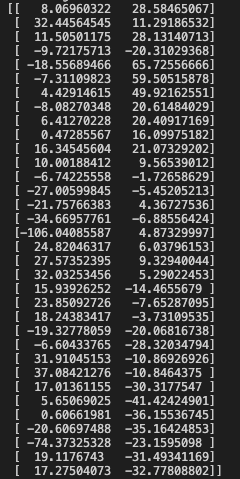
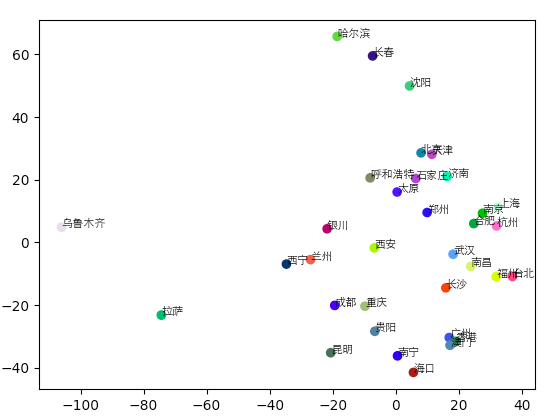
**模式识别第九周实验作业 许修为 2017010703**

1. 代码说明
   1. 本次实验使用python3.7，需要的第三方库有numpy，sklearn，pandas，xlrd和matplotlib。
   2. 第一个实验的代码为prml\_hw9\_Q1.py，需要和数据文件放在同一目录，直接运行程序，既可以得到绘制的图像。
   3. 第二个实验的代码为prml\_hw9\_Q2.py，需要和数据文件放在同一目录，可以直接运行。程序分两部分，第一问和第二问都有对应的代码块，注释中进行了标注。每次使用时去掉使用部分的注释，并给另一部分加上注释即可。
2. 城市距离的MDS可视化

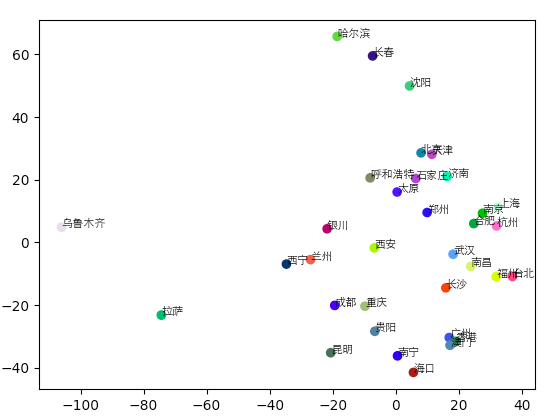
使用pandas读取excel，再使用MDS对距离矩阵进行降维（降到二维），得到每个城市的坐标，如下图：



画出这些点，得到城市的散点图，如下：

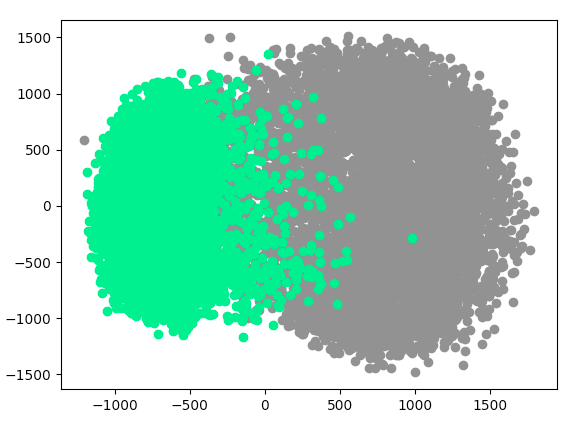


与真实地图相对比，发现总体上相对关系是非常相似的，但由于距离矩阵无法给出方位信息，因此上图与真实地图存在一定角度的偏差，若把上图顺时针旋转30度左右，如下图：

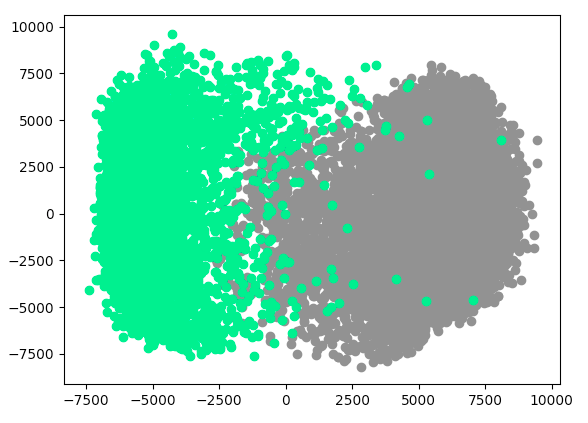


则可以看到，两者的城市分布是几乎相同的。

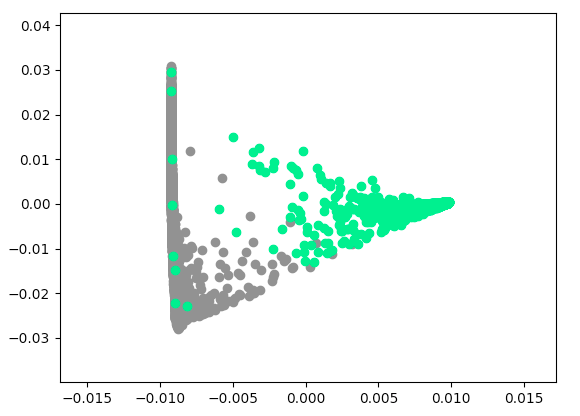
1. MNIST数据集的特征提取
   1. 使用PCA进行可视化，如图：



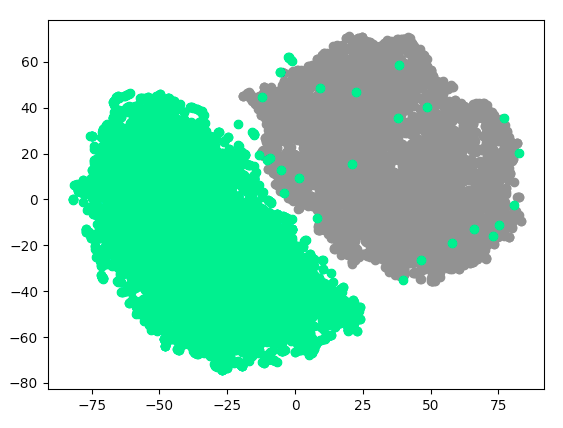
使用ISOMAP进行可视化，如图：



使用LLE进行可视化，如图：



使用tSNE进行可视化，如图：



从运行时间而言，PCA只需要不到一秒钟，而其他三种都超过了一分钟。

四种方法都能有效地将两类数据区分开，其中tSNE的效果最好。从可视化的结果还可以得知，进行降维后，这两类数据可以认为是线性可分的，用简易的线性分类器就可以很好地将训练集分开。

* 1. 将训练集和测试集数据一起进行PCA降维（原因将在下一问解释），使用逻辑斯蒂回归作为分类器，测试集准确率如下：

|  |  |
| --- | --- |
| 维度 | 测试正确率 |
| 1 | 0.3049 |
| 10 | 0.7882 |
| 20 | 0.8687 |
| 50 | 0.9038 |
| 100 | 0.9139 |
| 300 |  |
| 784 |  |

* 1. 应该先做PCA，再划分训练集和测试集。这样才能保证训练集和测试集仍具有相同的分布，由于PCA不使用标签信息，因此不会造成信息的泄露。