11.1 我们在第二章中曾经学习过 Fisher 判别方法，它将两类样本投影到类间方差尽可能大而类内方差尽可能小的方向上。这里我们考虑将 Fisher 判别进行推广，使其成为一种特征提取算法。

考虑K分类问题。我们尝试利用某种线性变换$W \in R^{D\times D'}$，将原始的$D(D>K)$维样本$x \in R^D$投影到$D'(D'<D)$维的空间中，投影结果记为$y \in R^{D'}$，则x与y的关系可表示为：

$$y = W^T x$$

通过以上投影方式，我们便能将原始D维的特征降维到D'维空间中，实现特征提取。

重新定义K分类问题的类内离散度矩阵与类间离散度矩阵：

$$S_W = \sum_{k=1}^{K}\sum_{n\in C_k}(x_n - m_k)(x_n - m_k)^T ; S_B = \sum_{k=1}^{K} N_k(m_k - m)(m_k - m)^T$$

其中，$m_k = \frac{1}{N_k}\sum_{n\in C_k} x_n$，$N_k$表示第$k$类样本的总数。

（1） 请证明类内离散度矩阵S_W与类间离散度矩阵S_B的和等于总离散度矩阵，即：

$$S_W + S_B = S_T = \sum_{n=1}^{N}(x_n - m)(x_n - m)^T$$

其中$\mathrm{m} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{1}{N}\sum_{k=1}^{K} N_k m_k$，$N$为样本总数。

（2） 请写出投影之后的样本在D'维空间中的类内离散度矩阵$\tilde{s}_W$与类间离散度矩阵$\tilde{s}_B$的表达式。

答案：

(1)

$$S_T = \sum_{n=1}^{N}(x_n - m)(x_n - m)^T = \sum_{k=1}^{K}\sum_{n\in C_k}(x_n - m)(x_n - m)^T$$

先考虑某个类别k

$$\begin{aligned}\sum_{n\in C_k}(x_n - m)(x_n - m)^T &= \sum_{n\in C_k}[x_n - m_k + m_k - m][x_n - m_k + m_k - m]^T \\ &= \sum_{n\in C_k}(x_n - m_k)(x_n - m_k)^T + 2\sum_{n\in C_k}(x_n - m_k)(m_k - m)^T \quad ① \\ &\quad + \sum_{n\in C_k}(m_k - m)(m_k - m)^T\end{aligned}$$

因为$\sum_{n\in C_k}(x_n - m_k) = 0$，所以①可转化为

$$\sum_{n\in C_k}(x_n - m)(x_n - m)^T = S_k + N_k(m_k - m)(m_k - m)^T$$

等式两边对类别求和，得

$$\sum_{k=1}^{K}\sum_{n\in C_k}(x_n-m)(x_n-m)^T=\sum_{k=1}^{K}S_k+\sum_{k=1}^{K}N_k(m_k-m)(m_k-m)^T \text{，即}$$

$$S_T=S_\omega+S_B \text{，其中 } S_B=N_k(m_k-m)(m_k-m)^T \qquad \text{证毕}$$

(2)

$$\tilde{m}_k=\frac{1}{N_k}\sum_{n\in C_k}\omega^T x_n=\omega^T m_k \qquad \text{同理 } \tilde{m}=\omega^T m$$

$$\begin{aligned}\tilde{S}_\omega&=\sum_{k=1}^{K}\sum_{n\in C_k}(y_n-\tilde{m}_k)(y_n-\tilde{m}_k)^T=\sum_{k=1}^{K}\sum_{n\in C_k}(\omega^T x_n-\omega^T m_k)(\omega^T x_n-\omega^T m_k)^T\\&=\omega^T S_\omega \omega\end{aligned}$$

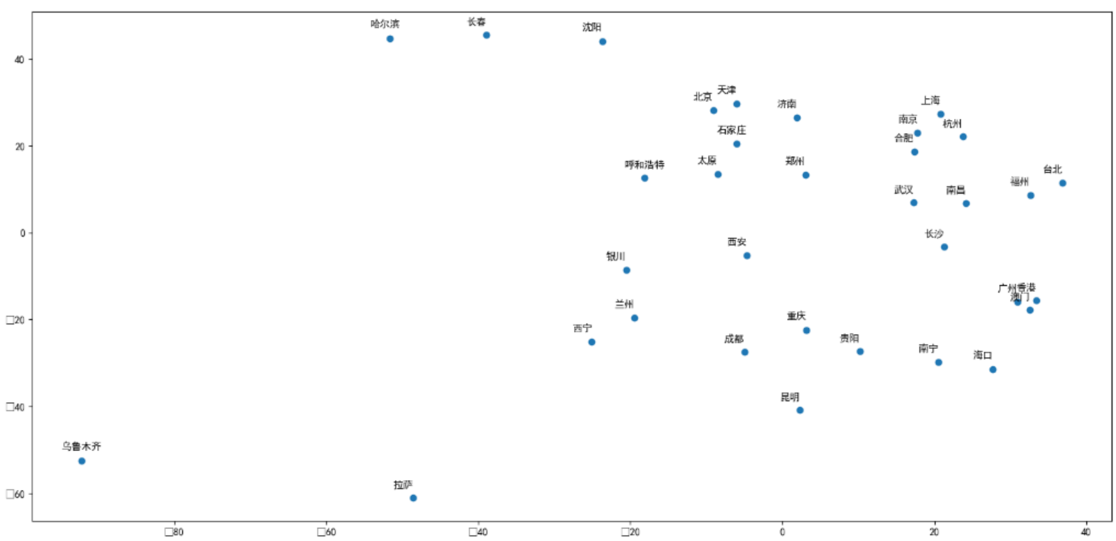
$$\begin{aligned}\tilde{S}_B&=\sum_{k=1}^{K}N_k(\tilde{m}_k-\tilde{m})(\tilde{m}_k-\tilde{m})^T=\sum_{k=1}^{K}N_k(\omega^T m_k-\omega^T m)(\omega^T m_k-\omega^T m)^T\\&=\omega^T S_B \omega\end{aligned}$$

11.2 计算机小实验 1：城市距离的 MDS 可视化

经典的 MDS（Multidimensional Scaling）方法起源于当我们仅能获取到物体之间的距离的时候，如何由此重构它的坐标。附件 city_dist.xlsx 中是 34 个城市之间的相对距离，请用 MDS 方法得到城市的二维表示并作图，简要分析你的可视化结果与真实地图上各个城市相对位置的差异。

答案：

可视化结果为：



可以看出，通过城市的相对距离，较为真实地还原了城市的各个位置，和

真实地图上的坐标很相似，比如哈尔滨、长春和沈阳，但也有个别城市的坐标和地图上稍有差异，比如太原和北京、乌鲁木齐和哈尔滨。

11.3 计算机小实验 2：MNIST 数据集的特征提取

在本题中，我们将数据集中的“0”和“9”两类样本集进行降维，并观察降维前后测试正确率的变化。

（1） 请分别用 PCA、tSNE 算法将训练集的数据降到二维并可视化，每一类样本用不同颜色的点表示，说明你从可视化图中能观察到什么信息。

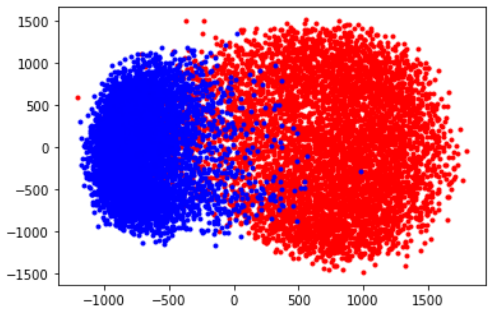
（2） 请用PCA算法将数据降维到1,10 ,50,300维，采用你认为合适的分类器分类，说明正确率随降维后维数的变化关系，并与不做降维之前的测试正确率进行比较。

（3） 请讨论对于分类问题，应该先做PCA降维再划分训练集、测试集进行学习；还是应该先划分训练集和测试集，再在训练集上做PCA。

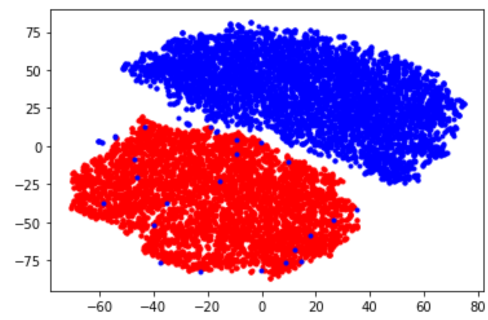
答案：

(1)

PCA 结果：



tSNE 结果：



从降维的结果可以看出，四种降维方法将数据降到2维后，数据的可分性还是较为满意的，其中PCA的效果相对而言两类样本的交叉更多，tSNE的降维效果

是相对较好的。

(2)

使用PCA降维方法对数据进行降维，并使用logistic回归进行分类。得到结果如下：

降维方法	1	10	20	50	100	300	未降维
PCA	0.9555	0.9872	0.9887	0.9933	0.9923	0.9872	0.9898

可以看到，未降维时分类准确率为 0.9898，在适当的维数(比如 50 维)，可以提高分类准确率(达到 0.9933)，而随着维数的继续升高，分类准确率又略有下降。

(3)

PCA 没有用到标签信息，不会造成信息泄露，应当使用所有数据进行 PCA 之后再进行训练集和测试集的划分