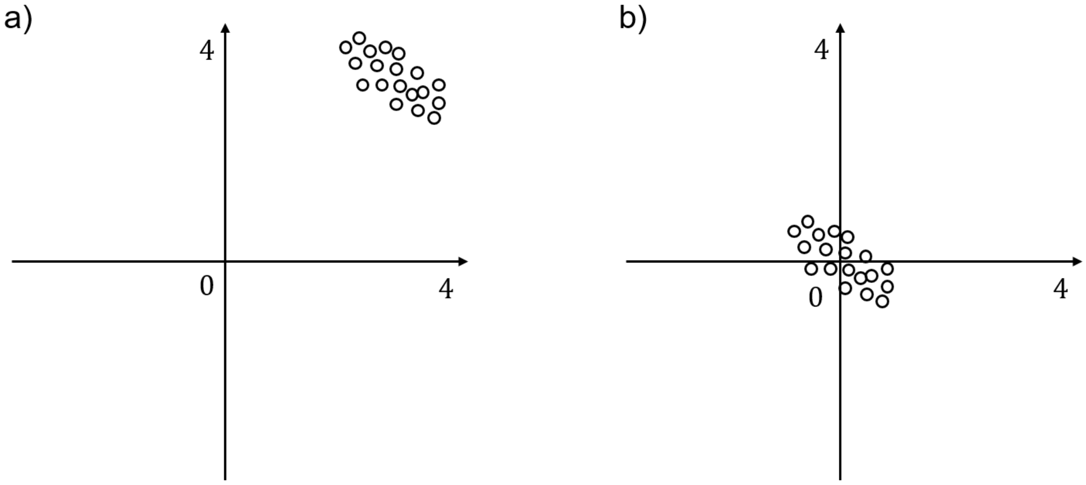
第十一章（特征提取与降维表示）作业

2020. 4. 14

11.1 下图（a）和（b）展示了样本点在二维平面内中心化（零均值化）前后的分布情况,请用直线在图中分别画出运行 PCA 算法得到的**第一主成分**的方向，并根据你所画的结果简述中心化（零均值化）对 PCA 算法的意义。



11.2 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：

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

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$m=\frac{1}{N}\sum_{n=1}^{N}x_n=\frac{1}{N}\sum_{k=1}^{K}N_k m_k$，$N$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。

（3） 类似投影到一维的 Fisher 判别，我们用如下判据来确定变换矩阵W：

$$J(W)=\frac{\prod_{diag}\tilde{s}_B}{\prod_{diag}\tilde{s}_W}$$

其中，$\prod_{diag}A$表示矩阵 A 的对角线元素的连乘。

请用 W,S_W,S_B 表示出$J(W)$，W 的每一列记为 $w_k\in R^D,k=1,2,\dots,D'$。并最大化$J(W)$，求解变换矩阵$W$。

11.3 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。

11.4 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“8”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、ISOMAP、LLE、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1,10,20,50,100,300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。