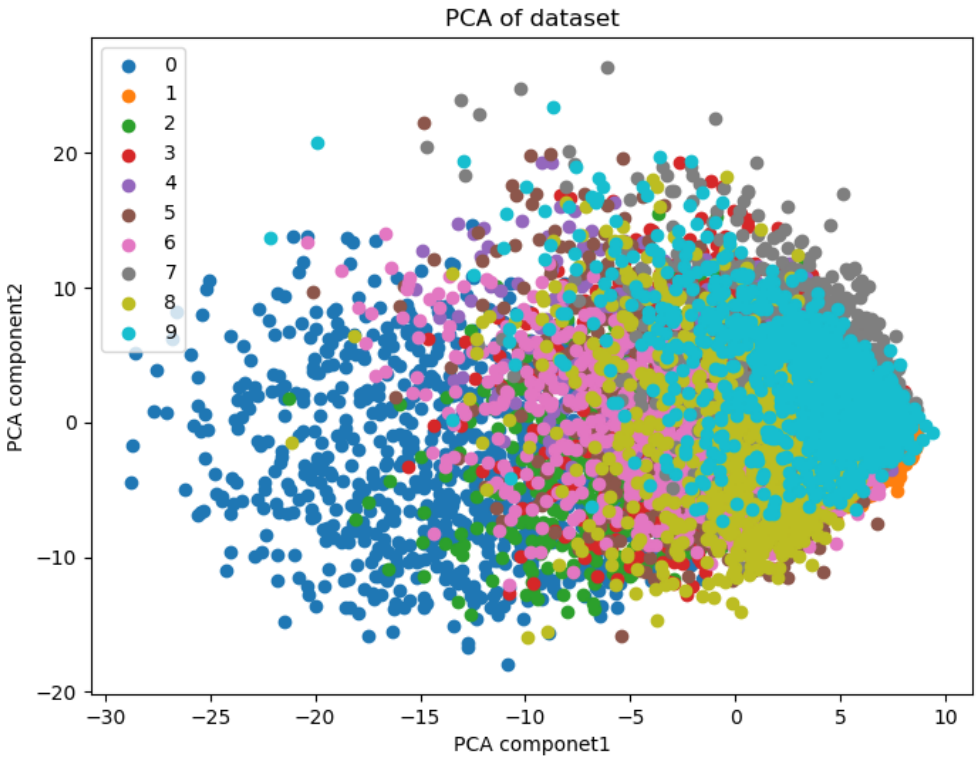
人工智能基础算法 第五次作业

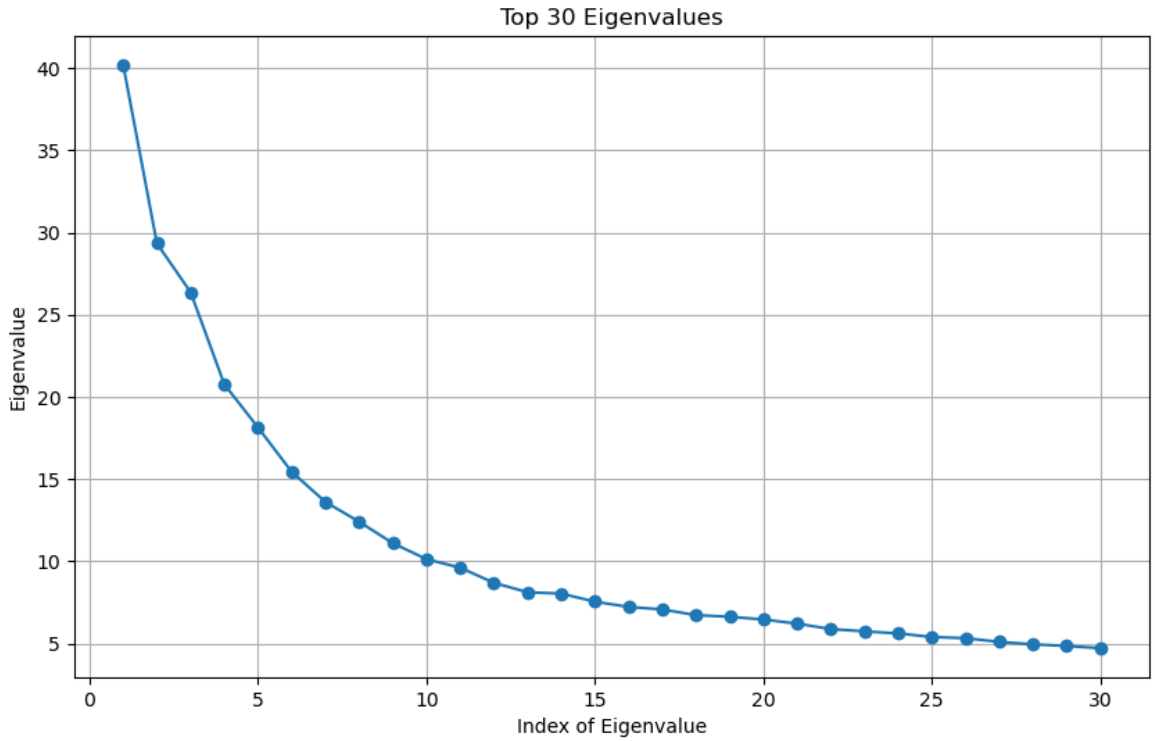
1

实现了PCA降维代码，主要包含在`pca`函数中，计算步骤为计算原样本数据的协方差矩阵、计算其特征值和特征向量、选择其前n个特征值向量、计算原有样本到这几个特征向量的投影，最终将投影依次作为其降维后的分量。本例中降维至二维。

本例计算时间为0.1862s，绘制数据降至二维的图像如下所示

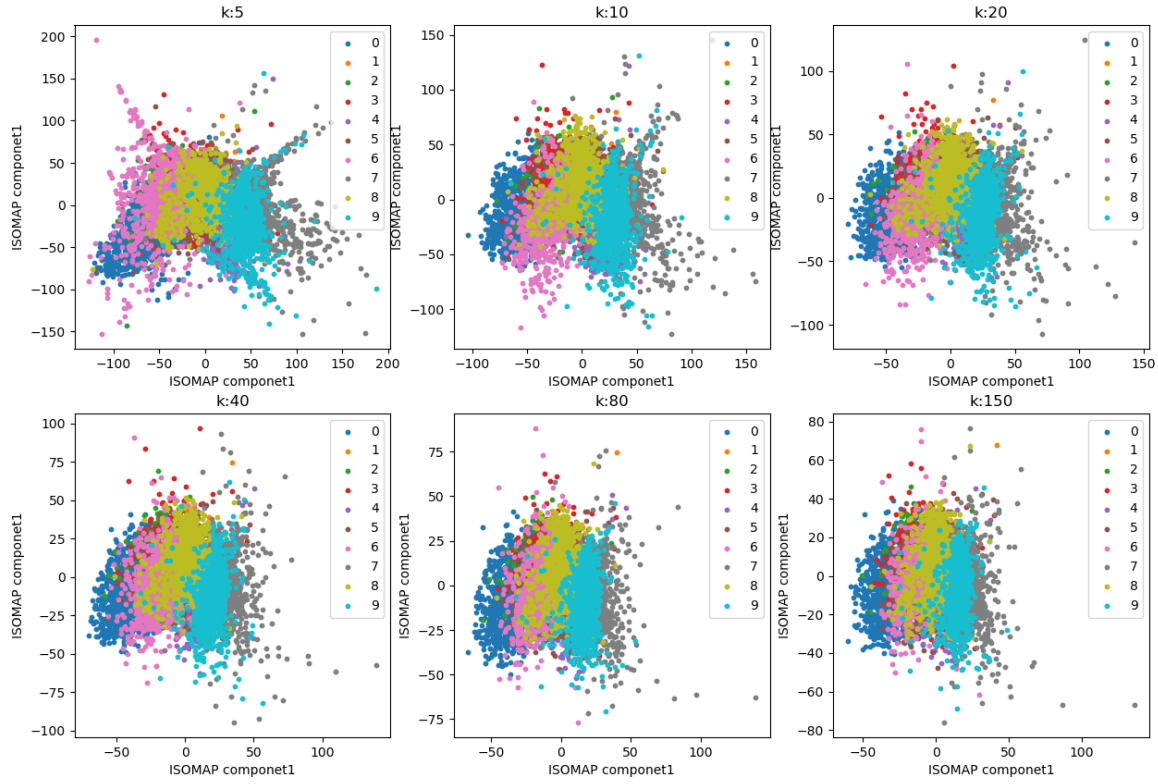


将协方差矩阵的特征值由大至小排列，绘制前30个特征值特征值曲线如下所示



2

使用ISOMAP算法（直接调用程序包），分别在最近邻数k=5,10,20,40,80,150六种情况下将数据降至二维，降维后的图像如下所示



计算总用时为468.60s，各k值的计算时间如下所示

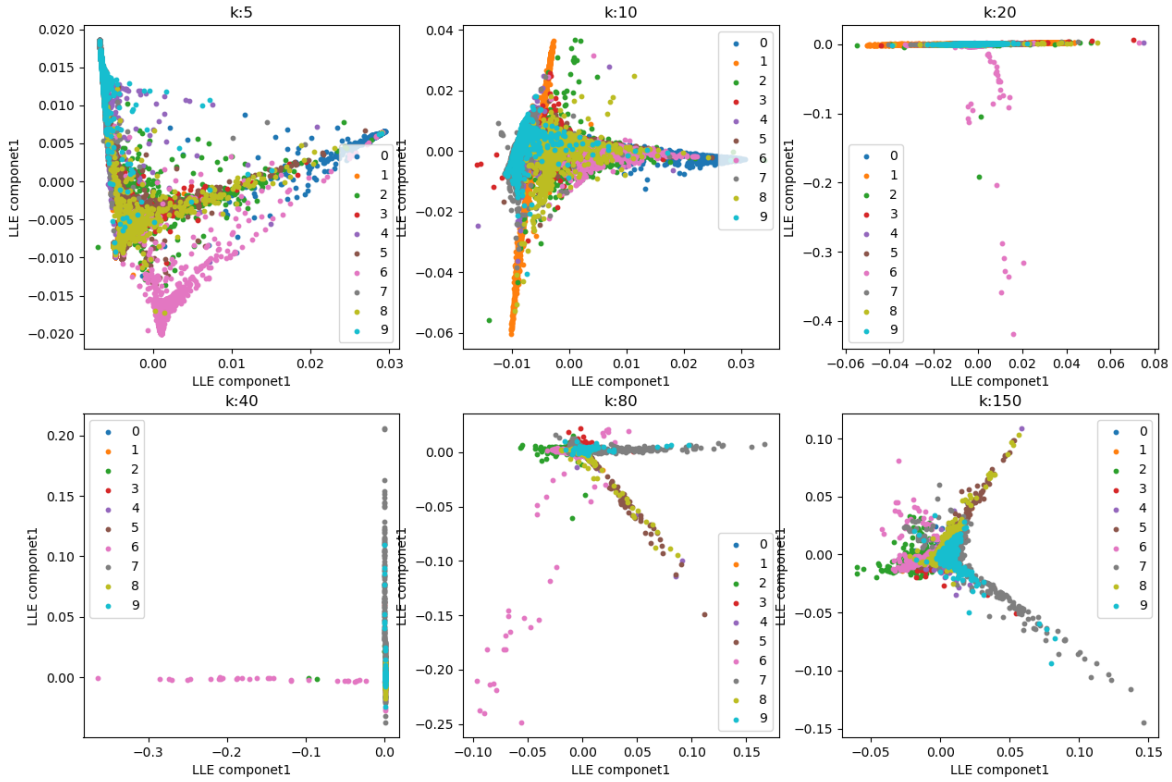
k	t/s
5	34.95
10	40.38
20	51.19
40	72.30
80	105.40
150	164.37

聚类效果可以从以下两个维度来进行判断：同一类的点的聚集程度和不同类的点之间的重叠程度，前者越集中、后者越小，则聚类效果越好，对样本的区分能力越强。

比较不同k值之间下的聚类效果可以发现，上述几种k值的选取基本都可以实现比较明显的聚类效果，而随着k的增大，同一类点的聚集程度明显变高，但同时，不同点之间的重叠程度也在变高，所有点明显变得更集中。综合以上两个维度进行考量，k=20的效果或许比较好。同时，随着k值的增大，计算时间逐渐增加，若综合上述进行考虑，选择k=20可能比较合适。

3

使用LLE算法（直接调用程序包），分别在最近邻数k=5,10,20,40,80,150六种情况下将数据降至二维，降维后的图像如下所示



计算总用时为307.14s，各k值的计算时间如下所示

k	t/s
5	7.87
10	20.83
20	42.37
40	52.25
80	69.03
150	114.77

观察图片可以发现，k=20、40的结果非常不好，几乎难以分辨不同类的点，不能实现聚，而k=80、150的结果虽然可以分辨不同类的点，但所有点集中在较小的区域内，不同类的点彼此重叠，聚类效果不佳；k=5、10的结果则可以较好地分辨不同类的点，聚类效果相对较好，其中，k=5的聚类效果更佳。再结合随k值不断增大的计算时间，认为k=5为最佳选择。

4

取一组q_i，$q_i \geq 0, i = 1, 2, \ldots, n$。令$Q_i = \frac{q_i}{\Sigma_{i=1}^{n} q_i}$，满足$0 \leq Q_i \leq 1$且$\Sigma_{i=1}^{n} Q_i = 1$。则求交叉熵关于$q_k$的导数

$$\frac{\partial}{\partial q_k}H(P,Q)$$
$$=\frac{\partial}{\partial q_k}(-\Sigma_{i=1}^{n}P_i(logq_i-log(\Sigma_{j=1}^{n}q_j)))$$
$$=\frac{1}{ln10}(\Sigma_{i=1}^{n}P_i\frac{1}{\Sigma_{j=1}^{n}q_j}-\frac{P_k}{q_k})$$
$$=\frac{1}{ln10}(\frac{1}{\Sigma_{j=1}^{n}q_j}-\frac{P_k}{q_k})$$
$$=\frac{1}{ln10}(\frac{q_k-P_k\Sigma_{j=1}^{n}q_j}{q_k\Sigma_{j=1}^{n}q_j})$$

要使$H(P,Q)$取最小值，则令$\frac{\partial}{\partial q_k}H(P,Q)=0$，则有

$$q_k-P_k\Sigma_{j=1}^{n}q_j=0$$

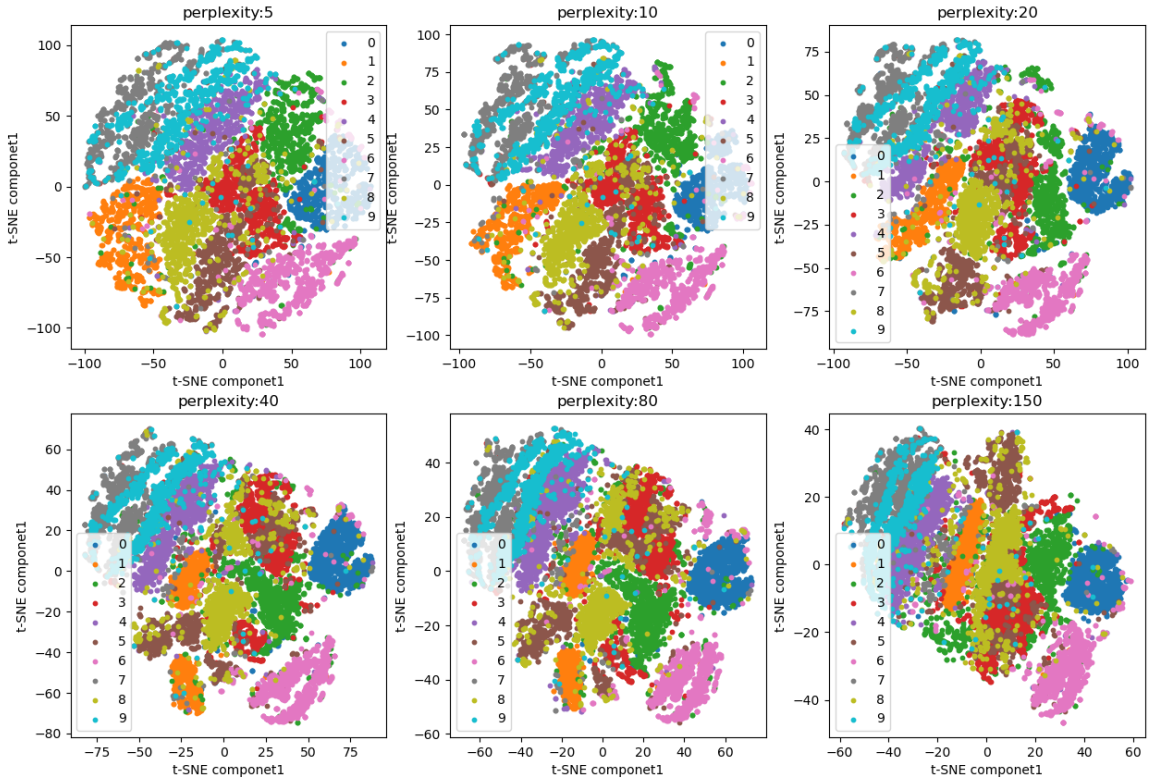
即

$$Q_k=\frac{q_i}{\Sigma_{j=1}^{n}q_j}=P_k\quad k=1,2,\ldots,n$$

即当$P=Q$时，$H(P,Q)$取最小值。

5

使用t-SNE算法（直接调用程序包），分别在困惑度参数$Perlexity$=5,10,20,40,80,150六种情况下将数据降至二维，降维后的图像如下所示



计算总用时为137.54s，各*Perlexity*值的计算时间如下所示

Perlexity	**t/s**
5	16.61
10	17.04
20	18.78
40	22.14
80	27.94
150	35.02

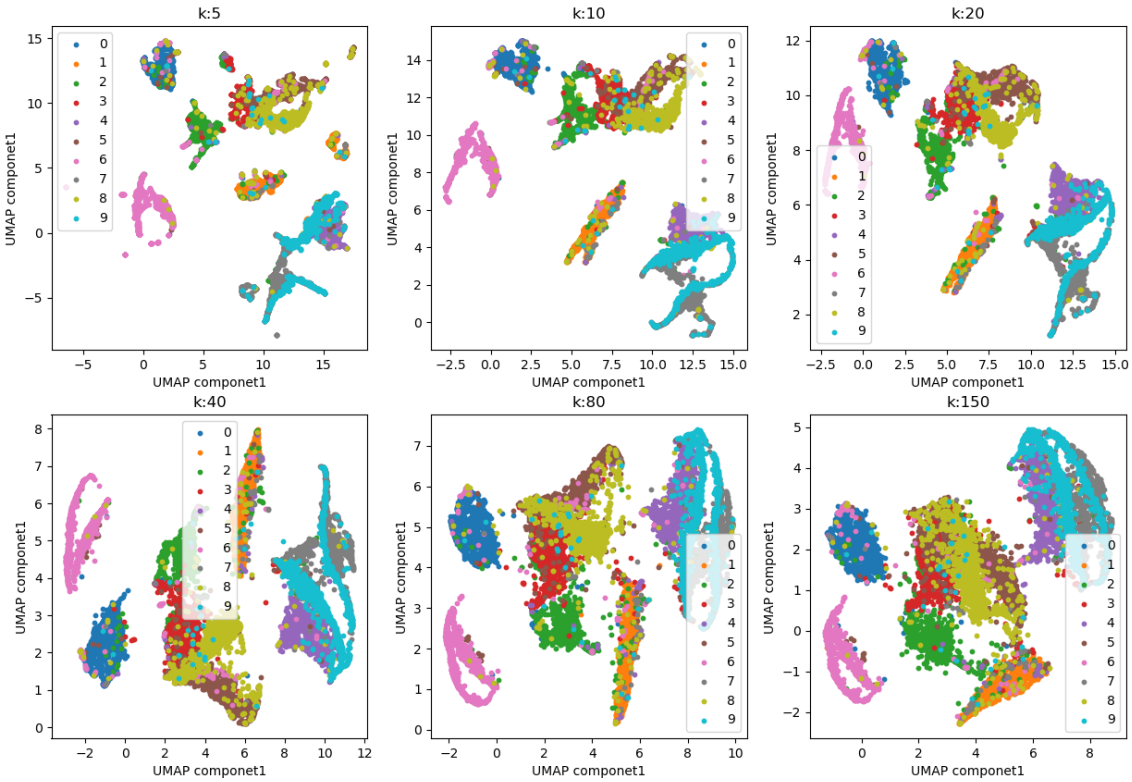
观察图片可以很明显地发现，相较于前两种方法，t-SNE算法的聚类效果明显更好。在不同的*Perlexity*值之间进行对比，可以发现*Perlexity*较小时图像更明显地将不同类的点划分在了图片的不同区域，其中k=5的聚类效果最好。再结合随k值不断增大的计算时间，认为k=5为最佳选择。

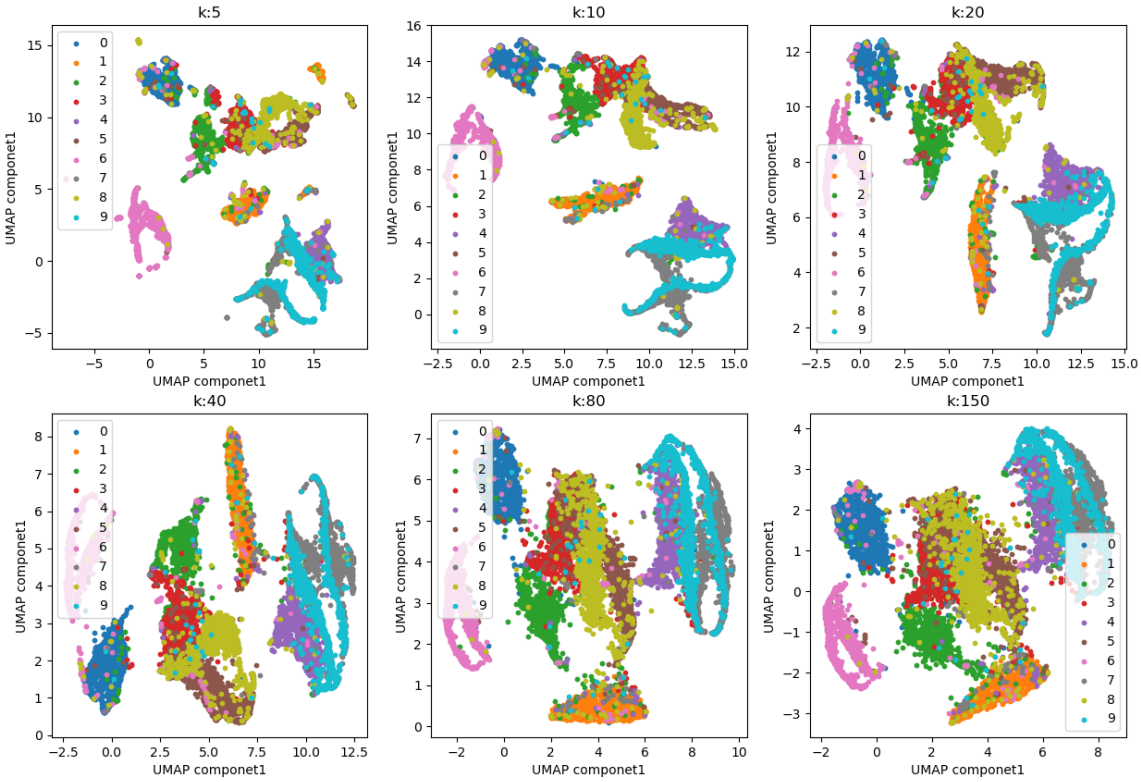
6

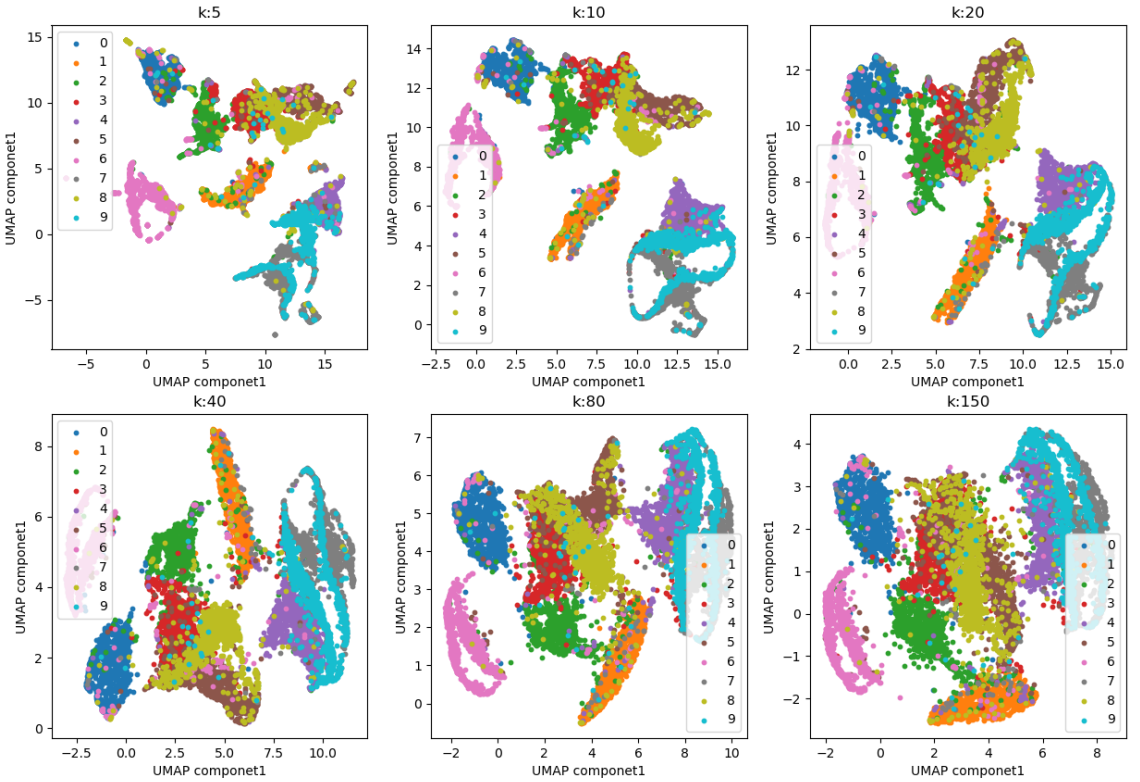
使用UMAP算法（直接调用程序包），分别在k=5,10,20,40,80,150六种情况下将数据降至二维，同时对参数`min-dist`进行选择。`min-dist`控制嵌入空间中点与点之间的最小距离，它决定了数据点在嵌入空间中的聚集程度。较小的 `min_dist` （例如 0.01 到 0.1）会使点更加聚集，适合于揭示数据的局部结构和小尺度的聚类，较大的 `min_dist` （例如 0.5 到 1.0）会使点更加分散，适合于揭示数据的全局结构和大尺度的聚类，0.1一般为默认值。

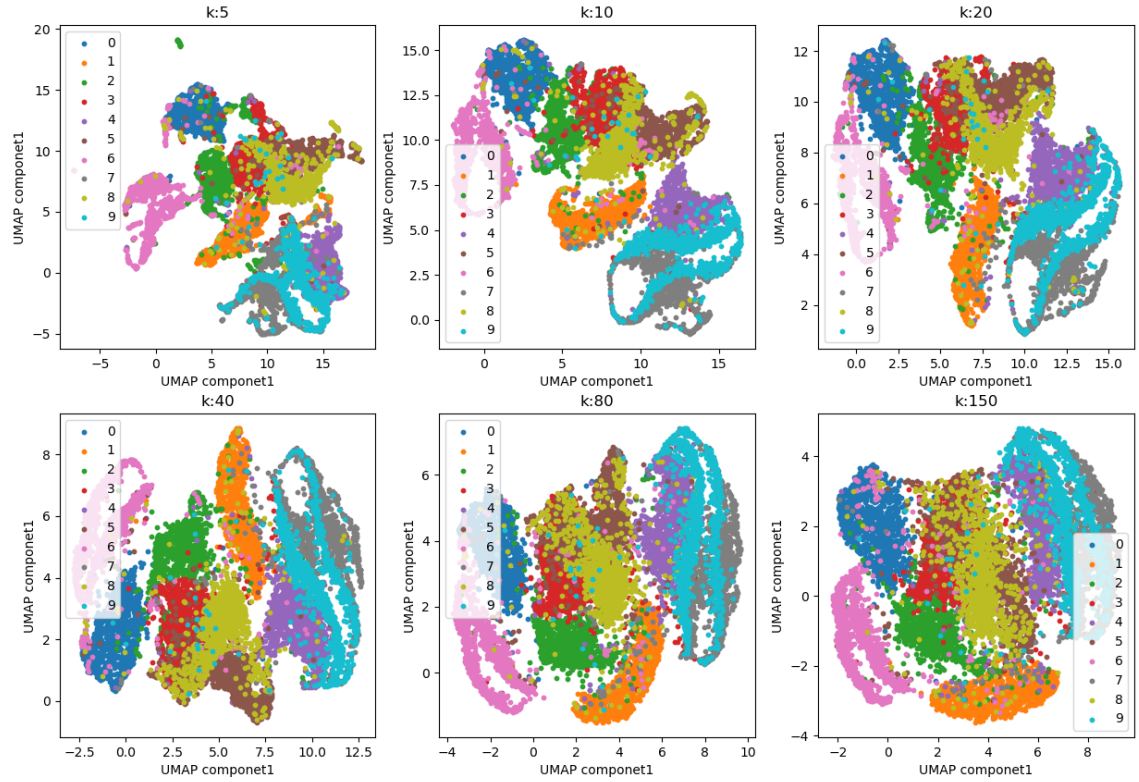
如果数据点的密度很高，则可以选择较小的`min-dist`值，以保留区域的细节；若数据点的密度较低，则可以选择较大的`min-dist`值，以避免过度聚集。

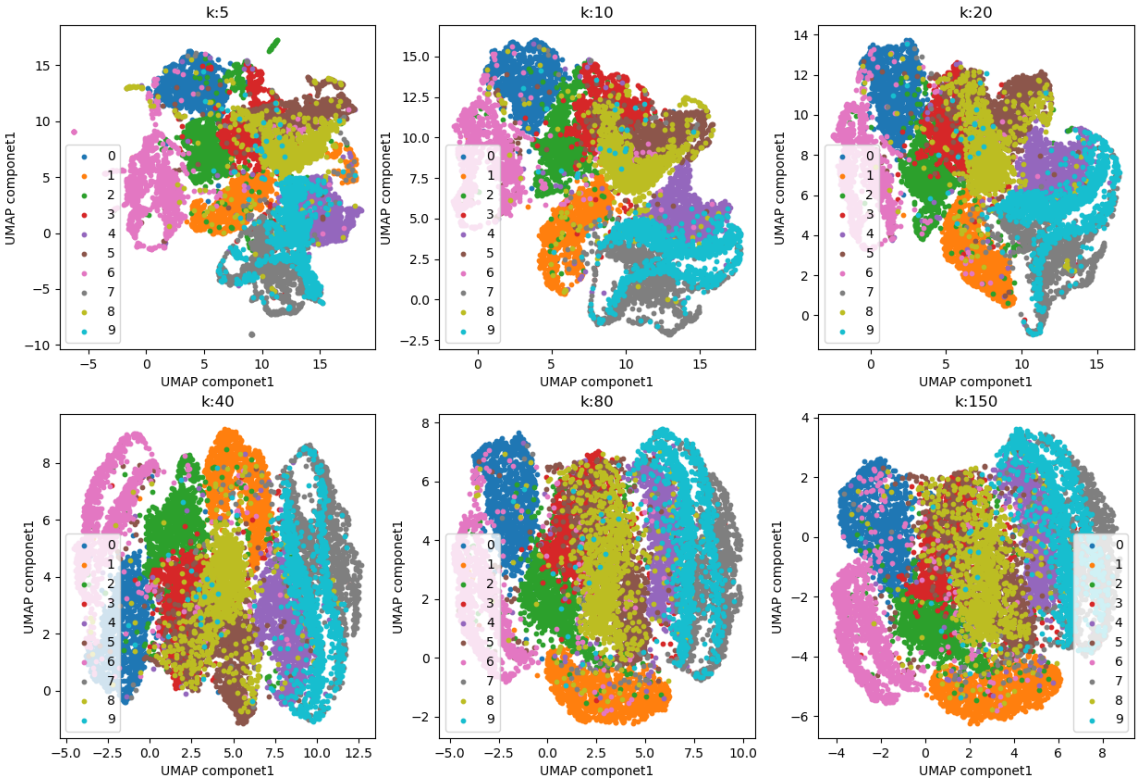
分别取min-dist=0.01,0.05,0.1,0.3,0.5，降维后的图像依次为




k:5
k:10
k:20
k:40
k:80
k:150
UMAP componet1
UMAP componet1
0
1
2
3
4
5
6
7
8
9


k:5
k:10
k:20
k:40
k:80
k:150
UMAP componet1
UMAP componet1
0
1
2
3
4
5
6
7
8
9


k:5
k:10
k:20
k:40
k:80
k:150
UMAP componet1
UMAP componet1
0
1
2
3
4
5
6
7
8
9


k:5
k:10
k:20
k:40
k:80
k:150
UMAP componet1
UMAP componet1
0
1
2
3
4
5
6
7
8
9

观察上述图像，本例中不同类的点间隔较远，数据密度较小，因而可以选择大一些的`min-dist`值，但过大的`min-dist`值仍可能导致数据过于重叠。综合考虑，我们选择min-dist=0.1。在此基础上，对不同k值的图像进行比较，认为k=80的聚类效果较好。

不同min-dist的计算时间基本相同，因而以默认值min-dist=0.1的计算时间为例进行说明。计算总用时为133.76s，各k值的计算时间如下所示

k	t/s
5	22.90
10	8.56
20	13.39
40	17.48
80	28.41
150	42.99

7

上述5种降维方法最显著的区别在于其将数据由高维转换至低维的过程中，所保持的信息不同。其中：

- PCA尽量保持数据的整体方差，最小化重构误差，主要保持数据在高维中的**全局结构**
- ISOMAP尽可能保持数据在高维空间的通过较近的点重构的距离，保持数据的**全局几何结构**
- LLE尽可能保持每个点在其邻域内与其他点的重构关系，保持数据的**局部结构**
- t-SNE保持每个点的邻居关系，同时保持**全局结构和局部结构**
- UMAP保持每个点的邻居关系，同时保持**全局结构和局部结构**

在此基础上，t-SNE和UMAP方法注重保持数据的局部相似性，兼顾了全局结构和局部结构，因而非常适用于对数据进行可视化，其进行降维可视化的效果非常好；LLE注重保持数据在邻域内的重构关系，聚焦在数据的局部结构上，相对忽略了全局结构，进行降维可视化所生成的图片仅保留了局部细节结构，损失了整体的全局结构，导致难以在整体上分辨不同类的数据，聚类效果不好，且它对噪声、邻域大小的选择非常敏感；ISOMAP则注重保持数据的全局几何结构，PCA则注重最小化重构误差、注重保持全局结构，二者均不太注重局部结构，因而所构建的可视化图像损失了较多的局部细节，同类的点在一定范围内聚成一团，聚类效果不佳。总结来说，t-SNE和UMAP的可视化效果较好。

而在计算时间方面，由于只需要计算协方差矩阵的特征值、特征向量，和各数据到对应特征向量的投影，PCA方法的计算量相对很小，计算速度非常快；而ISOMAP、LLE、t-SNE和UMAP相对其计算复杂度较高，计算时间较长。

（1）上述哪些方法重复实验会得到不同的结果？为什么？

- PCA、ISOMAP和LLE不会，其中PCA进行最小化重构方差计算，ISOMAP构建最短路径图和多维缩放（MDS），LLE保持邻域重构关系，均为确定性计算
- t-SNE和UMAP会，它们构建了一个基于概率关系的函数，并最小化该函数，其初始点的随机选取过程和可能存在的随机优化过程会导致结果具有随机性

（2） 降维得到的二维投影图中，哪些方法各类之间重叠部分较大，哪些较小？为什么？

- PCA、ISOMAP和LLE三种方法各类之间的重叠部分较大，这主要是因为PCA和ISOMAP相对忽略了数据的局部结构，导致各类数据分布缺少细节、在较大范围内聚成一团，不同类之间重叠较大，而LLE忽略了全局结构，不同类数据之间没有有效地分离开，仅保持了各自的局部结构，重叠部分较大
- t-SNE和UMAP重叠部分较小，这两种方法兼顾了全局结构和局部结构，对不同类数据进行了有效区分