人工智能基础算法 第五次作业

2024 年 11 月 6 日

题目内容：

本次作业仍然采用 MNIST 手写数字数据集，我们将用该数据进行降维可视化操作，我们的数据集中有 10000 个数据，其中每种数字各有 1000 个，示例代码中已经完成数据采样，请不要自己再去随机采样，保证大家使用相同的数据。

请使用 Python，Matlab，R 或其他编程语言完成任务并撰写简要报告，提交作业是要包含所有源码以及报告，可以补全提供的示例代码，也可以自己另写代码，本次作业除了第一题外，其余题目允许调用现有的程序包

作业要求：

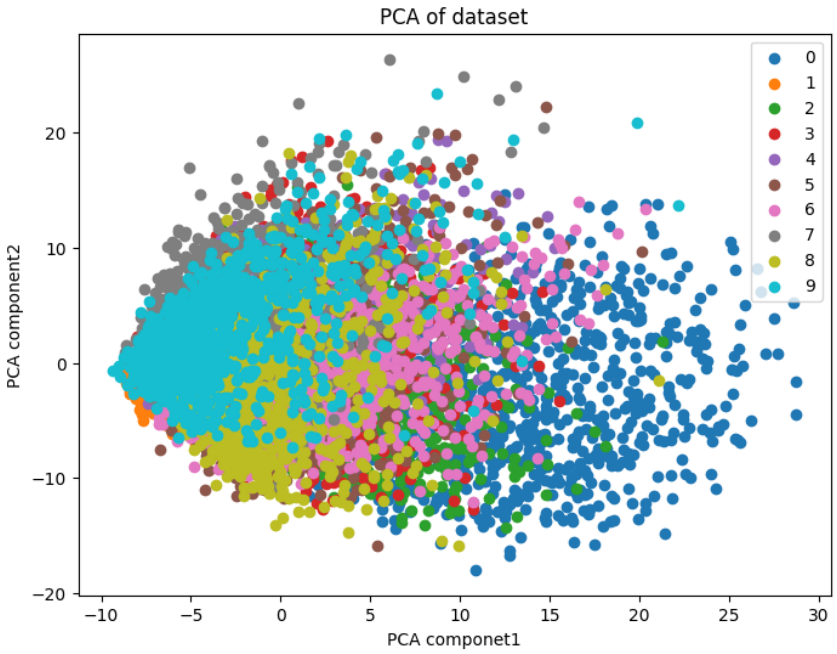
1.（15 分）PCA 降维。

（a）手写 PCA 降维代码，将原来的数据展成二维，并绘数据降到二维后的图像。要求：不同数字标签在图像中要用不同的颜色进行标注，并给出每种颜色代表哪个数字，实例代码中给出了参考样例，后面的绘图均需满足本要求。

（b）绘制特征值曲线，即横轴代表排序后特征值的位次，纵轴代表特征值的大小。绘制前 30 个特征值的曲线

注：可以使用现有的数学工具包，帮助进行 PCA 过程中的一些操作，如特征值分解等

示例图：



2.（15 分）ISOMAP 降维

分别给出在最近邻数$k = 5,10,20,40,80,150$六种情况下，使用 ISOMAP 降维到二维后的结果，要求绘制降维后的图像，并记录实验时间。本实验允许调用现有的程序包（如 python 中的 sklearn.manifold.Isomap），分析比较不同 k 下的聚类效果

3.（15 分）LLE 降维

分别给出在最近邻数$k = 5,10,20,40,80,150$六种情况下，使用 LLE 降维到二维后的结果，要求绘制降维后的图像，并记录实验时间。本实验允许调用现有的程序包（如 python 中 sklearn.manifold.LocallyLinearEmbedding），分析比较不同 k 下的聚类效果

4.（10 分）交叉熵

证明：定义两个离散概率分布：$P = \{P_1, \dots, P_n\}$和$Q = \{Q_1, \dots, Q_n\}$，我们可以定义这两个离散概率分布的交叉熵为：

$$H(P,Q) = -\Sigma_{i=0}^{n} P_i \log(Q_i)$$

假设P为给定的概率分布，Q为待优化的概率分布，则求证：

$P = Q$时，$H(P,Q)$取得最小值

注：本问题课上老师提供了一个思路，欢迎同学们互相讨论或者在微信群里讨论

5.（15 分）t-SNE 降维

分别给出$Perlexity = 5,10,20,40,80,150$六种情况下，使用 t-SNE 降到 2 维后的结果，要求绘制降维后的图像，并记录实验时间。本实验允许调用现有的程序包（如 python 中的 sklearn.manifold.TSNE）,其中$Perlexity$为困惑度参数，可以近似的理解为等效最近邻个数。

6.（15 分）UMAP 降维

分别给出$k = 5,10,20,40,80,150$六种情况下，使用 UMAP 降到 2 维后的结果，要求绘制降维后的图像，并记录实验时间。本实验允许调用现有的程序包（如 python 中的 umap）， 注：UMAP 还有另外一个参数 min-dist，请自行选取，报告参数数值，并说明理由。

7.(15 分)根据上面的实验结果，比较并讨论上面各种降维方法的区别，并回答下面的问题：

（1）上述哪些方法重复实验会得到不同的结果？为什么？

（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？

8（选做，不计分）对于基础 SNE 方法：

$$C = KL(P||Q) = \Sigma_i \Sigma_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

其中$q_{ij} = \frac{e^{-(y_i - y_j)^T (y_i - y_j)}}{\Sigma_{k \neq l} e^{-(y_i - y_j)^T (y_k - y_l)}}$

证明梯度可以被表示为：

$$\frac{\partial C}{\partial y_i} = 4\Sigma_j (p_{ij} - q_{ij})(y_i - y_j)$$

说明：

1. 作业附件中提供了 Python 样例代码，若使用 Python 完成，只在 homework5.ipynb 中添 加 TODO 内容即可。若使用其他语言，请在提交作业时删去文件中的样例文件。

2. 提交作业时请将代码及.pdf 格式的报告压缩为同一个.zip 文件，严禁抄

袭，否则本次 作业将会被记为 0 分!