人工智能基础算法 第四次作业

2024 年 10 月 24 日

题目内容：

之前我们已经使用了线性回归、人工神经网络等方法实现了对 MNIST 手写数字数据集的一些特征的识别，本次作业中我们将使用聚类方法来对手写数字进行分类。我们采用从 MNSIT 训练集中每个数字随机选取 300 个样本，总共 3000 个样本的聚类数据集。请使用 Python，Matlab，R 或其他编程语言完成任务并撰写简要报告，提交作业需包含所有源码以及报告。

本次作业所有聚类代码需要自己编写，不能调用现有聚类代码。在所有的问题中，可以放弃一个小问，不做回答，不扣分，但必须阐述放弃的原因（此原因供教师了解情况用）。

作业要求：

1. （15 分）实现 K 均值聚类算法，并以 K 作为用户可调参数。

2. （15 分）在给定的数据集上，实验 K=5，10，20，30，40，50 六种情形下聚类结果，在每种情形下，随机选取初始化簇中心 5 次。报告每次实验的 K 均值聚类费用函数值（见 LN6 第 21 页），计算时间，并可视化每个簇中心（可把每个簇中心向量转换成图像）。

3. （15 分）调研如何选取最佳 K 值，选取你认为合理的一个方案，阐述该方案，并报告基于问题 2 的最终结果。

4. （15 分）实现分级聚类算法，并以三种簇与簇之间距离定义（Single Link, Complete Link, 和 Average Link）作为用户可选项。

5. （15 分）在给定的数据集上，实验 Single Link, Complete Link, 和 Average Link 三种选项下聚类结果，报告计算时间，并画出如 LN6 第 38 页树状图。考虑到我们样本数很多，可以只画出从 1 簇到 100 簇的树状图。

6. （15 分）调研如何选取最佳层数或簇距离阈值，选取你认为合理的一个方案，阐述该方案，并基于问题 5 的最终结果，报告最终聚类结果的每个簇平均图像。

7. （10 分）比较并讨论 K 均值聚类和分级聚类算法，并以本次作业实验结果作为证据/支撑。

说明：

1. 作业附件 mnist_clustering_dataset.pkl 文件为从 MNSIT 训练集 0-9 中每个数字随机选取 300 个样本，总共 3000 个样本的聚类数据集。
2. 本次作业所有聚类代码需要自己编写，不能调用现有聚类代码。在所有的问题中，可以放弃一个小问，不做回答，不扣分，但必须阐述放弃的原因（此原因供教师了解情况用）。
3. 提交作业时请将全部代码及.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！