

《机器学习与数据挖掘》——作业 4

问题:

以 MNIST 数据集为例，探索 K-Means 和 GMM 这两种聚类算法的性能。数据下载 FTP 地址：

要求:

- 1) 使用 numpy, scipy 等科学计算包实现 K-Means 与 GMM 算法，在给定的 MNIST 数据集上训练，并在测试集上验证其性能。本次作业，直接将每张图片展开成 784 维，在 784 维的向量空间上进行聚类操作。本次作业不允许调用现成软件包（如 sklearn）来实现上述算法。
- 2) 在 K-Means 实验中，探索两种不同的质心初始化方法对聚类性能的影响。
- 3) 使用 EM 算法来训练 GMM 模型。在 GMM 实验中，探索使用不同结构的协方差矩阵（如：对角且元素值都相等、对角但对元素值不要求相等、普通矩阵等）对聚类性能的影响。同时，也观察不同初始化对最后结果的影响。
- 4) 使用聚类精度(Clustering Accuracy, ACC)作为聚类性能的评价指标。由于 MNIST 数据集有 10 类，故在实验中固定簇类数为 10。

实验报告需包含（但不限于）：

- 1) 简要描述 K-Means 和 GMM 的算法流程；
- 2) 采用的训练方法，包括参数初始化方法、优化方法以及其他的训练技巧等；
- 3) 通过观察实验结果，结合理论知识，比较 K-means 聚类方法和 EM 训练的 GMM 聚类方法之间的优劣；
- 4) 实验结果以及充分讨论，包括模型性能、训练时间、不同聚类算法的效果差异等。

将实验报告（.doc 或.pdf）和代码（不要数据）打包成一个文件， 文件包的命名规则为：
学号+姓名.tar，并上传到课程 FTP：

Due: 2021.5.23