线性分类器分类任务

概述

- 利用线性分类器对Kuzushiji-MNIST数据集中的测试集进行分类。

实验平台及数据说明

- Kuzushiji-MNIST是古日文的手写体识别数据集。该数据集由训练数据集和测试数据集两部分组成，其中训练数据集包含了60,000张样本图片及其对应标签，每张图片由28×28的像素点构成；训练数据集包含了10,000张样本图片及其对应标签，每张图片由28×28的像素点构成。

任务说明

- 任务一：对Kuzushiji-MNIST数据集进行预处理，然后在处理后的训练集上学习一个多类线性分类器，并对处理后的测试集进行分类。
- 任务二：利用 PCA 降维方法对 Kuzushiji-MNIST 数据集进行降维，然后在降维后的数据上完成多类线性分类器的训练和测试。要求比较应用 PCA 降维技术前后，分类器准确率的变化。（对于降维后的数据，可以尝试利用可视化方法展示结果。）

Tips

- 推荐语言：Matlab、Python（可采用Numpy，Pandas，Matplotlib等基础代码集成库）、C++。
- 不得使用集成度较高，函数调用式的代码库（如Python环境下的sklearn，PyTorch，Tensorflow等）。
- Kuzushiji-MNIST数据集以二进制形式保存，需要编写读写二进制数据的程序完成对图片、标记信息的初步提取。
- 多类线性分类器的实现可以考虑一对一、一对其余等策略。
- Kuzushiji-MNIST数据集最初来源于参考文献[1]，对其兴趣的同学可以在课余时间进一步阅读文献[1]。

作业提交格式要求

- 提交两份测试集分类结果文件（原始测试数据“t10k-images.idx3-ubyte”在分类器上的分类结果和降维后的测试数据“t10k-images.idx3-ubyte”在分类器上的分类结果），请分别命名为**task1_test_prediction.csv**和**task2_test_prediction.csv**，文件格式参照sample_submission.csv。
- 需提供完整的**代码文件、预处理完的数据文件和测试集分类结果文件**，将以上内容打包压缩，**压缩文件命名格式：学号-姓名-线性分类器分类任务实验。**
- 尽量以相对路径的形式索引数据集，便于我们对代码进行复现。

- **成果若有雷同，一律按0分处理。**

参考文献

[1] Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep learning for classical japanese literature. CoRR, abs/1812.01718, 2018.