

# 线性分类器分类任务

## 概述

- 利用线性分类器对MNIST数据集中的测试集进行分类。

## 数据说明

- MNIST是著名的手写体数字识别数据集。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 60000 张样本图片及其对应标签，每张图片由 $28 \times 28$ 的像素点构成；训练数据集包含了 10000 张样本图片及其对应标签，每张图片由 $28 \times 28$ 的像素点构成。

## 任务说明

- **任务一**：对MNIST数据集做恰当**预处理**，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**LDA降维技术**进行对MNIST数据集的降维工作，在降维后的数据集上完成多类线性分类器的训练和测试。要求比较应用LDA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）。
- 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。
- MNIST数据集最初来源于参考文献<sup>[1]</sup>，对其刚兴趣的同学可以在课余时间进一步阅读文献<sup>[1]</sup>。

[1] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

## 作业提交格式要求

- 提交两份测试集分类结果文件（一份降维前的分类器在“t10k-images.idx3-ubyte”上的分类结果，一份降维后的分类器在“t10k-images.idx3-ubyte”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。
- **代码和报告若有雷同，一律按0分处理。**