朴素贝叶斯分类实验

概述

- 利用朴素贝叶斯算法，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28×28的像素点构成；测试数据集包含了 10000 张样本图片及其对应标签，每张图片由28×28的像素点构成。该数据集的训练数据可以通过东大云盘 https://pan.seu.edu.cn:443/link/E2457AD6A5B3CC111AD45E90F0A8030D 下载，也可以通过 http://yann.lecun.com/exdb/mnist/ 地址下的下载链接下载。

实验内容

- 在课程学习中同学们已经学习了贝叶斯分类理论并掌握了其基本原理，即利用贝叶斯公式$p(\omega_j|x) = \frac{p(x|\omega_j)p(\omega_j)}{p(x)}$，对$p(\omega_j|x)$作出预测，由于$p(x)$为一固定值，所以一般不在计算过程中求得$p(x)$的具体值。在实际运用中，为了方便计算，通常假设数据特征之间相互独立，即$p(x|\omega_j) = p(x_1|\omega_j) \cdot p(x_2|\omega_j) \ldots \cdot p(x_d|\omega_j),\ x \in \mathbb{R}^d$，这便是著名的朴素贝叶斯算法。

- MNIST 数据集本身以二进制形式保存，所以首先需要选择合适的编程语言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工作。读取了图片信息后，发现每个像素点的值在[0,1]区间中，这是图像压缩后的结果，所以可以先将像素值乘以 255 再取整，得到每一个点的灰度值。将图像二值化，得到可以用于分类的28×28个特征向量以及对应的标签数据，之后便可以交由贝叶斯分类器进行学习。

- 基于 MindSpore 平台实现算法，对相同的数据集进行训练，并与不使用 MindSpore 平台实现的算法对比结果（包括但不限于准确率、算法迭代收敛次数等指标），并分析结果中出现差异的可能原因，给出使用 MindSpore 的心得和建议。

- （加分项）使用 MindSpore、sklearn 等平台提供的相似任务数据集（例如，其他的分类任务数据集）测试自己独立实现的算法，并与 MindSpore、sklearn 等平台上的官方实现算法进行对比，进一步分析差异及其成因。

实验要求

- 推荐使用 Python（在独立实现算法时，可采用 Numpy，Pandas，Matplotlib 等基础代码集成库；在使用 MindSpore 平台时，可使用平台提供的代码集成库）。
- 在独立实现算法时，不得使用集成度较高、函数调用式的代码库（如 sklearn，PyTorch，Tensorflow 等）。
- 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。

实验报告格式

- 需要提供完整的可运行代码文件、测试集分类结果文件和实验报告，将以上内容打包

压缩，压缩文件命名格式：学号-姓名-xxx 实验。实验报告和代码注释应尽量详细。需要以相对路径的形式索引数据集或文件，便于我们对代码进行复现。

- 实验报告内容参照报告模板，包括问题描述、实现步骤与流程、实验结果与分析、实验的心得体会（谈谈你自己的实现和 MindSpore 实现的差异、你在使用 MindSpore 平台过程中遇到的问题，以及想对平台改进提出的建议）、一个总的心得体会（谈一谈你对这门课程理论及实验的感悟与体会）。
- 代码和报告若有雷同，一律按 0 分处理。
- 若存在疑问，可以联系：pr_seu@163.com