

KNN 分类任务

概述

- 利用 KNN 算法对 Iris 鸢尾花数据集中的测试集进行分类。

数据说明

- 鸢尾花数据集内包含的 3 类分别为山鸢尾 (Iris-setosa)、变色鸢尾 (Iris-versicolor) 和维吉尼亚鸢尾 (Iris-virginica)，共 150 条记录，每类各 50 个数据，每条记录都有 4 项特征：花萼长度、花萼宽度、花瓣长度、花瓣宽度。标签 0、1、2 分别表示山鸢尾、变色鸢尾、维吉尼亚鸢尾。
- 数据集已被划分为训练集、验证集和测试集，分别存储于 data 文件夹下的 train.csv, val.csv 和 test_data.csv 文件。其中 train.csv 和 val.csv 包括 data 和 label 字段，分别存储着特征 $X \in \mathbb{R}^{N \times d}$ 和标记 $Y \in \mathbb{R}^{N \times 1}$ ， N 是样例数量， $d = 4$ 为特征维度，每个样例的标记 $y \in \{0, 1, 2\}$ 。test_data.csv 仅包含 data 字段。

实验内容

- **任务一：**利用欧式距离作为 KNN 算法的度量函数，对**测试集**进行分类。实验报告中，要求在验证集上分析近邻数 K 对 KNN 算法分类精度的影响。
- **任务二：**利用马氏距离作为 KNN 算法的度量函数，对**测试集**进行分类。在马氏距离中， M 为半正定矩阵，正交基 A 使得 $M = AA^T$ 成立。给定以下目标函数，在**训练集**上利用梯度下降法对马氏距离进行学习：

$$f(A) = \min_A 1 - \sum_{i=1}^N \sum_{j \in \Omega_i} p_{ij}$$

其中， Ω_i 表示与 x_i 属于相同类别的样本的下标的集合， p_{ij} 定义为：

$$p_{ij} = \begin{cases} \frac{\exp(-d_M(x_i, x_j)^2)}{\sum_{k \neq i} \exp(-d_M(x_i, x_k)^2)} & j \neq i \\ 0 & j = i \end{cases}$$

d_M 为：

$$d_M(x_i, x_j) = \|Ax_i - Ax_j\|_2$$

实验中，矩阵 A 的维度 e 可任意设置为一合适值，例如 $e = 2$ 。

实验报告中请对优化过程的梯度计算公式进行推导，即给出 $\frac{\partial f}{\partial A}$ 的计算公式

- **任务三：**基于 MindSpore 平台提供的官方模型库，对相同的数据集进行训练，并与自己独立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。

- **加分项：**使用 MindSpore 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 平台上的官方实现算法进行对比，并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）。

实验报告格式

- 以相对路径的形式索引数据集，便于我们对代码进行复现。
- 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-KNN 分类任务。实验报告和代码注释应尽量详细。
- 提交测试集预测结果文件时，请注意任务一、任务二、任务三各需提交一个预测结果文件，并命名为 task1_test_prediction.csv, task2_test_prediction.csv, task3_test_prediction.csv，文件格式参照 sample.csv，便于对实验结果进行评估。
- 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、每个实验的心得体会（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）。
- 代码若有雷同，一律按 0 分处理。
- 若存在疑问，可以联系：casie_che@163.com