

KNN分类任务

概述

- 利用KNN算法对wine数据集中的测试集进行分类。

数据说明

- wine葡萄酒数据集是UCI上的公开数据集。数据集包含由三种不同葡萄酿造的葡萄酒，通过化学分析确定了葡萄酒中含有的13种成分的含量。数据集的相关信息如表1所示：

表1 wine数据集相关信息

样例数量	特征维度	特征类型	类别数量
178	13	数值	3

- 数据集已被划分为训练集、验证集和测试集，分别存储于data文件夹中的train_data.mat, val_data.mat, test_data.mat。train_data.mat和val_data.mat文件包含data, label字段，分别存储着特征 $X \in \mathbb{R}^{N \times d}$ 和标记 $Y \in \mathbb{R}^{N \times 1}$ 。其中，N是样例数量， $d = 13$ 为特征维度，每个样例的标记 $y \in \{1, 2, 3\}$ 。test_data.mat文件仅包含data字段。
- 训练集中包含部分特征值缺失的样例。

任务说明

- **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。马氏距离是一种可学习的度量函数，定义如下：

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T M (x_i - x_j)}$$

其中， $M \in \mathbb{R}^{d \times d}$ 是一个半正定矩阵，是可以学习的参数。由于M的半正定性，可将上述定义表述为：

$$d_M(x_i, x_j) = \sqrt{(x_i - x_j)^T A^T A (x_i - x_j)} = \|Ax_i - Ax_j\|_2$$

其中，矩阵 $A \in \mathbb{R}^{e \times d}$ 。故，马氏距离可以理解为对原始特征进行线性映射，然后计算欧式距离。

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：

$$f(A) = \underset{A}{argmax} \sum_{i=1}^N \sum_{j \in C_i} p_{ij}$$

其中， C_i 表示与样例 x_i 同类的样例集合， p_{ij} 定义为：

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

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。

Tips

- 推荐语言：Python（可采用Numpy, Pandas, Matplotlib等基础代码集成库）、Matlab、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）。
- 建议考虑对数据进行必要的预处理，以应对特征值缺失等问题。

作业提交格式要求

- 需提供完整的**代码文件**和**测试集预测结果文件**，将以上内容打包压缩，**压缩文件命名格式：学号-姓名-KNN分类任务实验**。
- 提交测试集预测结果文件时，请注意任务一、任务二各需提交一个预测结果文件，并命名为**task1_test_prediction.csv, task2_test_prediction.csv**，文件格式参照sample_submission.csv。
- 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。
- **代码若有雷同，一律按0分处理。**