《机器学习基础 2022 秋》东南大学微电子学院

Assignment 1（截稿时间 10 月 10 日 24 点）

以 MNIST 数据集实现 Fisher Linear Discriminant（FLD）的分类以及降维功能。20 分题。

构造以下文件结构

---根目录 （自定义）

---DataSet //将数据集解压至此文件夹，保留原文件名。

利用 python-mnist 库读取 mnist 原始二进制文件。从二进制数据集文件 train-images-idx3-ubyte 中提取相应的数组。其中包含数字 0~9，每个数字约 6000 张的共计 60000 张图片。每张图片的尺寸是 28*28，每个像素是 int32 的整形数，表示该像素的灰度值，数值取值范围 0~255。参考文件: 测试数据集读取.ipynb。

将作业命名为：作业 1_姓名.ipynb，该作业完成以下任务。

任务一： 采用 Fisher Linear Discriminant 方法对数字“5”与数字“8”进行分类实验。即

步骤 1：预处理数据集。

1.1）从 minist training set 中提取数字“5”与数字“8”的样本，各取任意 1000 张，其中 800 张组成新的训练集，剩余 200 张组成测试集。

1.2）将测试集以及训练集中每张图片的像素值归一化到 0~1 之间， 方法是：像素值除以 255。再将每张图片调整成 784 维特征列向量，记为 X_i。至此，获得数字“5”与数字“8”的 training set，各自 800 个样本；以及它们的 testing set，各自 200 个样本。并画出数字“5”的 training set 以及 testing set 的均值图像。

步骤 2：使用 training set，构造 FLD 二分类器，在 784 维向量空间中寻找最佳投影方向 W，将数字“5”与“8”的样本最大程度清楚地区分。

2.1）根据《机器学习.周志华（第一版）》page61 页，公式 3.33 计算类内散度矩阵 S_w，公式 3.34 计算类间散度矩阵 S_b。FLD 待定参数 W 的方向由公式 3.39 确定。请注意，通过奇异值分解求取 S_w 矩阵的逆，以保持数值稳定性。

2.2）将 training set 中数字“5”“8”各自 800 个样本投影到 W 方向，并绘制它们投影后内积 W^TX_i 的分布。该直方图的横坐标表示 W^TX_i 取值范围，纵坐标表示概率。采用 numpy 中的 histogram（）构造各自的直方图，并利用 python 的 Matplotlib 库或者 Pandas 库可视化，将两类样本投影后数据分布绘制在同一张图上，寻找交点的横坐标作为阈值 thre。

2.3）确定阈值之后，在 testing set 中测试数字“5”“8”各自 200 个样本，并评估 error rate，给出 confusion matrix。Error rate 的计算方法是错分样本个数除以 testing set 总样本个数。

任务一完毕。

任务二： 使用 FLD 的降维功能，在 784 维特征向量空间寻找三个投影方向，分别是 W_1，W_2，W_3，将数字“5”“8”在这三个互相正交的方向所确定的特征空间内最大程度、清楚地区分。并在剩余数字类中，通过类间散度、类内散度的评估，寻找能与“5”“8”最大程度区分的数字是几。请给出论证过程。

步骤一：预处理数据集。

从 minist training set 中提取数字“0”到“9”各自任意 5000 个样本。同任务一，将每张样本图片的像素值归一化到 0~1 之间，再调整成 784 维特征列向量，记为 X_i。

步骤二：利用《机器学习.周志华（第一版）》page62 的公式 3.42，计算各类数字样本的类内散度 S_{wi}。

步骤三：取数字“5”“8”的 5000 个预处理之后的特征列向量样本，利用公式 3.41 计算 S_w，公式 3.43 计算 S_b，并构造 $S_w^{-1}S_b$。取该矩阵最大三个非零特征值所对应的的特征向量，即为 W_1，W_2，W_3。建议使用 svd 分解求 $S_w^{-1}S_b$ 的特征值，可使用 np.linalg.svd 函数。

步骤四：

4.1）将剩余 8 种数字类样本投影到由 W_1，W_2，W_3 所决定的三维空间。要求绘制出三维空间内各类数字样本的投影图，以不同的颜色区分数字。

4.2）在该空间内通过这些数字类与“5”“8”样本之间类间散度（公式 3.43）、类内散度（公式 3.41）的评估，取数字类为横轴，以 S_b/S_w【注 1】为纵轴，绘制曲线，并确定能与“5”“8”最大程度区分的数字。

任务二完毕。

注 1：三轴分别计算类内方差、类间距。三轴的 Sb 之和做分子，Sw 之和做分母，得到 S_b/S_w。