LAB2：无监督学习 - MNIST聚类实验

在本实验中，需要你来具体实现PCA降维和高斯混合模型（GMM）聚类，具体实验内容见 `MNIST.md`文件。

截止日期

11.31 23:59 (UTC+8)

迟交1/2/4/7天将扣除10%/20%/40%/60%分数

实验评分

本实验满分100，具体组成为：

- 代码补全(40pt)
- 实验效果(30pt)
- 报告(30pt)

环境配置

本课程将使用一个连续的环境配置，你可以在此开一个新的环境，之后实验中我们都将以这个环境做基础，如果需要，会在基础上安装更新更多的python库。

本实验激活虚拟环境以及安装必须的库

```
conda activate ai25
pip install -r requirements.txt
```

报告要求

你的报告应该符合规范，包含以下内容：

- 描述你的实验流程，包括数据处理、模型训练、结果可视化的完整过程(4pt)
- 说明你使用的降维方法（PCA或AutoEncoder）和高斯混合模型（GMM）聚类及参数设置以及调整过程(5pt)
- 展示关键的可视化结果，报告最好的聚类和生成结果 (输出的图片)，并且分析不同降维方法的展示效果(4pt)
- 对于MNIST.pdf中问题的回答(15pt)
- 课程反馈：本次实验你花费的时间是多少？（必填）(2pt) 任何课程/实验/作业的建议（可选）

推荐同学们使用Latex来编写报告

实验效果评分

- Code(40%):见 MNIST.pdf
- Performance(30%):

  我们在原始数据空间 (784 维) 中使用 `davies_bouldin_score`度量聚类的性能，DB score（戴维森堡分数）衡量聚类簇间的分离性和簇内的紧密性，**值越小表示聚类效果越好**。评分时，会将你生成的聚类标签与原始测试数据输入评分脚本，计算 DB score 并排行。

  具体流程如下：

  1. 使用你的训练好的PCA对测试集特征进行降维。
  2. 利用你的训练好的GMM模型对降维后的特征进行聚类，得到每个样本的聚类标签。
  3. 在原始784维特征空间下，结合聚类标签，使用 `sklearn.metrics.davies_bouldin_score` 计算测试集的聚类得分。

$$DB_{yours} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\sum_{i=1}^{n} d(x_i, \mu_k)}{\sum_{j=1}^{n} d(x_j, \mu_j)} - 1 \right)$$

具体评分公式待定~

- Report(30%):
  - 记录实验流程(4%)
  - 记录你调试超参数的过程(5%)
  - 报告最好的聚类和生成结果 (输出的图片)(4%)
  - 回答问题(15%):见 MNIST.pdf
  - 反馈(2%):见 MNIST.pdf

提交格式

你的提交应该包含以下文件：

```
├── submission.py          # 你实现的核心代码（PCA和GMM）
├── report.pdf             # 实验报告
└── results/               # 实验结果目录（可选，调试用）
    └── [时间戳]/
        ├── config.yaml
        ├── pca.npz
        └── gmm/
```

提交方式：

将文件打包为zip格式，命名为 `<学号>-<姓名>-LAB2.zip`，提交到BB系统

例如：`PB123456-张三-LAB2.zip`