作业十五：Johnson–Lindenstrauss 降维效果实测

任务 & 数据集

本作业的总体目标是：给定 d 维欧氏空间下 n 个点，使用 Johnson–Lindenstrauss 降维并测量降维误差。本题使用的是数据来自于 ANN-Benchmark[1]的 Fashion-MNIST 数据集。该数据集有 768 维，原本有 60000 个点，但我们作业的数据仅从中随机抽取了 1000 个点。作业需要用的数据在压缩包内的 data.txt 中。

数据集格式 data.txt 文件中的数据总共有 $n = 1000$ 行。每行有 $d = 768$ 个整数，用来描述一个点的坐标。

具体任务描述 你的任务是使用课上讲的 JL 算法将这 n 个点降到 m 维，并对不同的 m 测试最大误差和平均误差。具体来说，我们需要考虑一个 $m \times d$ 维的随机矩阵 G，其中每一项都是独立同分布的标准正态变量 $N(0,1)$，然后定义映射 $f(x) := Gx$。设对于两个点 x, y，$\mathrm{dist}(x,y) := \|x-y\|$ 定义为它们的欧氏距离。对于 $1 \le i < j \le n$，定义数据点 x_i, x_j 上的降维误差为

$$\mathrm{err}(i,j) := \frac{|\,\mathrm{dist}(f(x_i), f(x_j)) - \mathrm{dist}(x_i, x_j)|}{\mathrm{dist}(x_i, x_j)}.$$

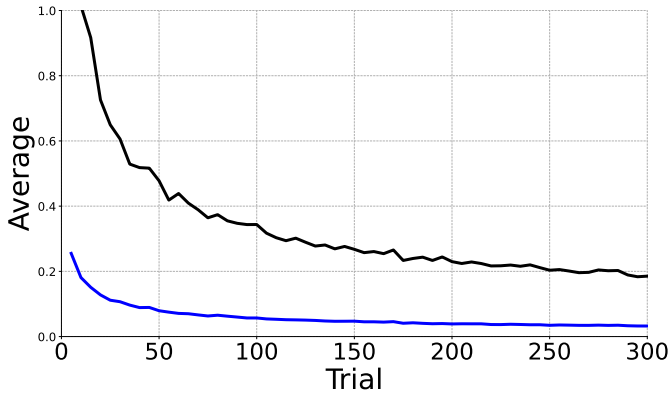
定义降维在数据集上的误差的最大值 $\epsilon_{\max}^{(m)}$ 和平均值 $\hat{\epsilon}^{(m)}$ 为：

$$\epsilon_{\max} := \max_{1 \le i < j \le n} \mathrm{err}(i,j),$$
$$\hat{\epsilon} := \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathrm{err}(i,j).$$

请对 $m = 5, 10, 15, \ldots, 300$ 分别计算 $\epsilon_{\max}$ 和 $\hat{\epsilon}$，并利用随附的 draw.py 绘制横轴是 m，纵轴分别是 $\epsilon_{\max}$ 和 $\hat{\epsilon}$ 的两条曲线。为了结果的稳定性，你可能需要对每个 m 都进行多次随机试验（即采样不同的 G），并对 $\epsilon_{\max}$ 和 $\hat{\epsilon}$ 取平均值后绘图。图 1是一个格式示例。

[1] https://github.com/erikbern/ann-benchmarks

图 1: 参考格式



提交作业要求

你需要提交一个压缩包 (zip 格式) 包含以下内容：

- 代码文件 (C++ 语言)。**你的代码需要能够输出你所提交的趋势图中的数据。**
- 图片文件：可使用随附的 draw.py 来生成。draw.py 使用了 Python 的matplotlib绘图库。你当然也可以使用其他工具绘图，但需要与参考格式类似的格式。

作业提交 将代码和图片集成在一个压缩包中提交。压缩包需命名为“HW15 + 学号 + 姓名”，并提交至教学网。

分数组成

提交的压缩包中**必须包含代码以及图片，同时代码的输出与趋势图相匹配，否则本题得 0 分。**各部分得分占比如下：

- 代码：用于查重，只要是自己写的即可，严禁抄袭。分数占 70%
- 趋势图：变化趋势合理即可，分数占 30%