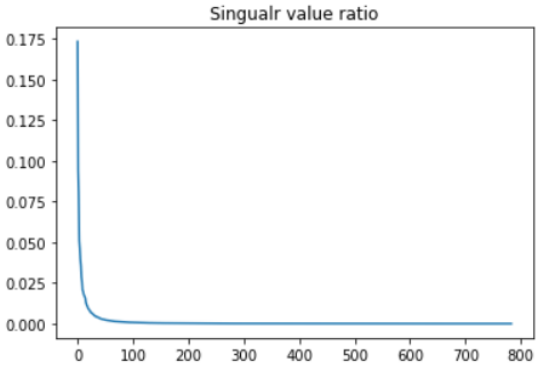
Ex9 record and report

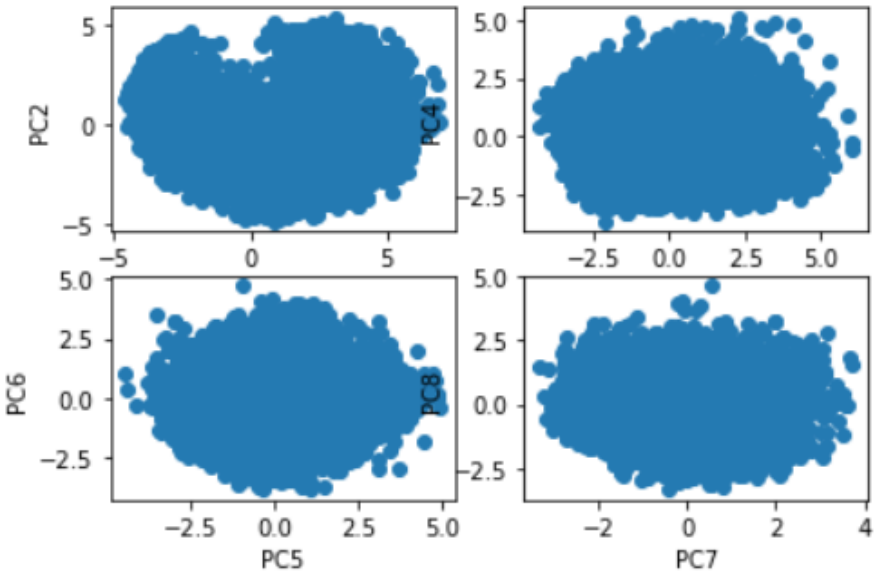
王敏行 id：2018012386 wangmx18@mails.tsinghua.edu.cn

本次实验数据为 MNIST 手写体数据为“7”和“1”的数据。

流程方面，先利用 torchvision 加载数据，并从中挑选出标签为“7”的数据。将 1*28*28 的黑白图片（0，1 二值张量）转换成 1*784 维的数据，以供下游分析。对数据进行 PCA 降维，并记录所有特征值/奇异值的占比，绘制下图。可见前几十个特征的奇异值占了大部分总奇异值，故取前 20 维进行后续的分析。也尝试了二维的 t-SNE 降维，效果不佳，见 notebook 输出。

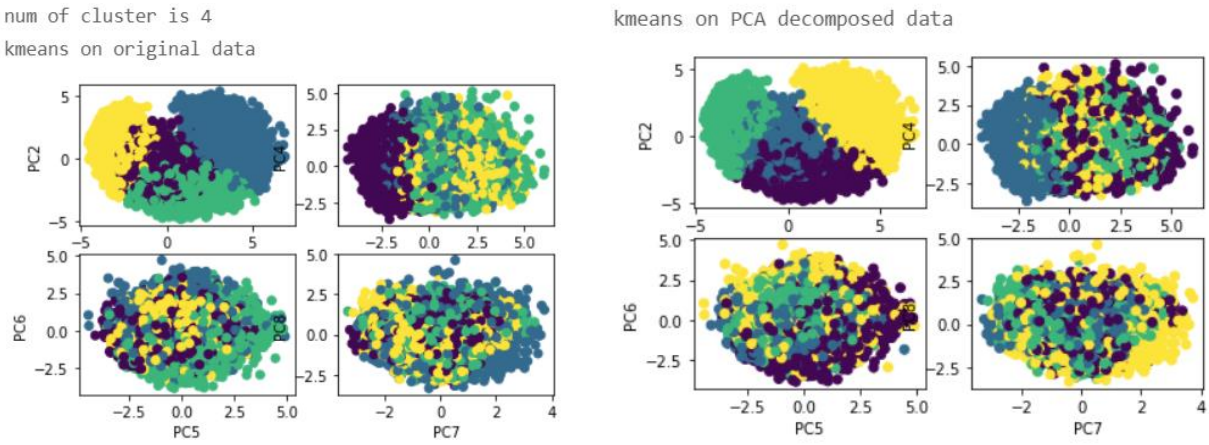


挑出前八个主成分绘图如下：



对前 20 个主成分组成的数据进行 k-means 聚类分析，并将分类的结果绘制在主成分变换的空间中。聚落的数量分别尝试 2~10，结果见 notebook 的输出。

肉眼判断，4 个聚落时分类效果较好，结果如下：



可见，低微数据的聚类结果与原始数据在前 4 个维度上差别不明显，分类界限较为清晰。

同样对数字“0”进行上述分析，发现其 t-SNE 结果可分性更好。用前 10 个主成分进行 k-means 分类，分类的结果在 t-SNE 图如下所示。