作业要求：

- 每人完成总分值为 100 分的候选题目
- 使用的方法模型不限、编程语言不限
- **要求提交物：**
  - 实验报告：应包括任务定义、输入输出、方法描述、结果分析（性能评价）、编程和实验的软硬件环境、其它需要说明的项
  - 代码：除开源工具以外的其它源码或可执行文件
- **提交方式：**每次作业布置后一周内，通过电子邮件发送至课程邮箱 yuansassignment@163.com，邮件主题为：学号-姓名-作业编号

其它说明：

- 关于分组：
  - 不采用多人分组，每人独立完成至少 100 分值的作业
- 关于加分：
  - 如果对于一个题目提供了不同的解决方案，或在一个解决方案之上提供了改进方案，则可额外加最多 10 分，具体根据完成情况确定
  - 最后一次课为作业演示时间，演示者通过 PPT 向大家介绍自己的某一个或几个作业，演示者则可额外加最多 10 分，具体根据演示情况确定
  - 然而，无论如何，总分不超过 100 分
- 诚信说明：经鉴定为抄袭或被抄袭，两种情况均得 0 分
- 每人最高得分为 100 分

Problem assignment

题目 1：聚类

我们已经学习过无监督学习、K-means 和 GMM。本题目要求对 cluster.dat 进行聚类。cluster.dat 包含了若干二维输入数据（但不包含其输出）。

方法 1. K-means (20 points)
使用 K-means 模型进行聚类。尝试使用不同的类别个数 K，并分析聚类结果。
附加题 (BONUS: 10 points)
按照 8:2 的比例，随机将数据划分为训练集和测试集。至少尝试 3 个不同的 K 值，并画出不同 K 下的聚类结果，及不同模型在训练集和测试集上的损失。对结果进行讨论，发现能解释数据的最好的 K 值。

方法 2. Gaussian Mixture Model (30 points)
使用 GMM 及 EM 算法。尝试使用不同个数的混合成分。对不同的高斯分布，尝试使用关联的协方差矩阵和独立的协方差矩阵。分析聚类结果。
附加题 (BONUS: 10 points)
按照 8:2 的比例，随机将数据分为训练集和测试集。对于不同个数的混合成分，绘制随着迭代的进行，模型在训练集和测试集上的似然，并对结果进行讨论。

题目 2：特征降维和特征学习

MINST 是一个手写数字数据集，包括了若干手写数字体及其对应的数字，共 60000 个训练样本，10000 个测试样本。每个手写数字体被表示为一个 28*28 的向量。

方法1. PCA or kernel PCA (30 points)
使用 PCA 或 kernel PCA 对数据进行降维。观察前 2 个特征向量所对应的图像，即将数据嵌入到 R^2 空间。绘制降维后的数据，并分析 2 维特征是否能够足以完成对输入的分类。

方法2. Autoencoder (30 points)
使用自动编码器学习输入的特征表示。尝试设计一个全链接前馈神经网络。尝试使用不同的损失函数和正则化方法。
附加题 (BONUS: 10 points)
模型训练中，你可以尝试任何可以提升模型性能的合理的方法。例如其它的网络结构、设计多个隐藏层、引入降噪自动编码器等任何你能想到的方法。计算模型在训练集和测试集上的损失，并对结果进行讨论。