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
  - 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
- 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

Task 1: Clustering

In class, you have been introduced to the unsupervised learning and the K-means algorithm ang GMM. Our goal is to categorize the two-dimensional dataset cluster.dat into several clusters.

Method 1. K-means (20 points)

Implement K-means method on the cluster.dat. You should try different numbers of clusters.

Do something extra! (BONUS: 10 points)

Split the dataset using 80-20 train-test ratio. Train your predictor using newly-implemented K_means function. Iterate over k, for each report the training and testing loss. Plot training and testing loss versus k. Plot the samples for three choices of k. Pick k which reveals the structure of the data. Comment the results.

Method 2. Gaussian Mixture Model (30 points)

Implement EM fitting of a mixture of gaussians on the cluster.dat. You should try different numbers of mixtures, as well as tied vs. separate covariance matrices for each gaussian.

Do something extra! (BONUS: 10 points)

Split the dataset using 80-20 train-test ratio. Plot likelihood on training and testing vs iteration for different numbers of mixtures.

Task 2: Dimensionality Reduction and Feature Learning

In dataset speech.json, you are given the text of 3119 speeches made by presidential candidates between 1996 and 2016. Each speech was either made by a Democrat or a Republican. The goals of this assignment are as follows:

Method 1: PCA or kernel PCA (30 points)

Implement principal components analysis on this data. Use PCA to find 2-dimensional compressed features. These compressed features allow you to embed the speeches into R^2. Plot the resulting embedded data points. Do you think these 2 features are enough to classify the speeches?

Method 2: Autoencoder (30 points)

You will now perform a neural network with autoencoder constraints. You can design a fully-connected classifier. You should try experiments with different loss functions and regularizations.

Do something extra! (BONUS: 10 points)

In the process of training your network, you should feel free to implement anything that you want to get better performance. You can modify the solver, implement additional layers, use different network architecture, use an ensemble of models, or anything else that comes to mind.