本次作业需要用到的数据文件：

- 数据集：dataset3.txt　含 954 个样本（469 女、485 男）

作业 2016-11-27. 用身高体重等数据进行聚类的实验

1. 把数据 **dataset3.txt** 作为未知样本集(即忽略每个样本的类别信息),用 **PCA** 对 **dataset3.txt** 进行降维，画出各个主成分上的方差，根据方差分布确定选取几个主成分来构成数据新的表示。

2. 从以下两题中任性一题完成进行聚类分析（也鼓励两题都做）：

2a) 在以上得到的主成分表示上，用 **C** 均值方法对全部样本进行聚类分析，试验 **C=2、3、4、5、6** 几种情况，连同 **C=1** 的情况一起画出 **C** 个聚类的误差平方和随类别数变化的折线，观察能否用非监督学习方法发现样本中有意义的聚类。选择其中一个最可能有意义的 **C***。将聚类结果以适当的方式显示出来，对聚类结果进行分析和讨论。

2b) 分别用原始的十维特征和 **PCA** 选出的主成分做分级聚类，考查采用不同特征、不同距离度量选项对结果的影响，讨论将数据聚为几类更合理。将聚类结果以适当的方式显示出来，对聚类结果进行分析和讨论。

作业要求：

1、 交作业日期：2016 年 12 月 12 日前（含）打包提交到网络学堂。

2、 提交内容：

a) 实验报告（PDF 文件）

- 题目、姓名、学号、班级、日期
- 实验方法
  - 实验设计、采用的方法、原理、过程和程序来源说明
- 结果与讨论

b) 结果数据

- 以适当的格式提交最后所选用的聚类结果文件，须对文件格式和数据的含义约定进行详细的说明。
- 程序代码及说明、程序运行步骤及参数。

3、 关于编程和讨论：

鼓励自己编写程序（用任何语言），也允许使用工具包，甚至不禁止使用他人程序。在实验报告及程序报告中须明确写明程序出处和作者。如采用他人程序，程序报告中“程序代码及说明”直接使用原作者的版本（须注明），但“程序运行步骤及参数”则根据自己的实验自己完成。

对实验结果的分析，鼓励同学间讨论，但实验和报告必须独立完成。

如发现抄袭或未经说明的引用，本次作业将记-10 分。如无法区分抄袭者和被抄袭者，则都记-10 分。**如发现捏造数据，本次作业记-20 分。**