数据挖掘与大数据分析
Assignment 5

1. 数据集（15 分，每个数据集 5 分）

下载三个数据集（带有类别标签，但聚类过程并不使用类别标签，类别标签用于计算评价指标），下载以后，仔细阅读数据集的使用说明，理解其用途及每一列数据的含义。

2. 聚类分析（60 分）

- 方法
  - （20 分）编程实现 DPC 算法；
  - （40 分）调用 scikit-learn 包中的 K-Means、DBSCAN、SpectralClustering（谱聚类）、EM 算法（高斯混合模型）；
- 评价指标

  自行选择两个评价指标，可选的范围为：NMI、RI、purity、Silhouette Coefficient（自行思考为何不用 Accuracy）。

使用上述聚类算法分别对下载的三个数据集进行聚类分析，获取聚类结果及评价指标的值。

3. 撰写技术报告（25 分）

以科技论文的形式撰写 assignment 的技术报告，呈现实验结果并进行分析。

- 自行设计实验，达到以下目的
  - 对比各个聚类算法在同样数据集上的聚类质量，分析哪个算法为何在哪个数据集上能取得较好的聚类结果。
- 实验部分应对数据集进行介绍，参考文献中给出该数据集的原始出处，并在报告正文中第一次出现该数据集的地方添加对文献的引用；
- 对实验结果的呈现，必须以文字形式进行阐述、解释或者说明，不能只是简单地展示结果的图，否则会减分；调整图的大小，使之清晰美观，否则会减分；

- 报告应以正规的书面语言进行客观的阐述，切勿使用口语化的表达方式或使用随意的网络用语；
- 插图应使用矢量图，图、表要添加编号与标题，并在正文中引用其编号；
- 报告中对使用的算法应引用其出处的参考文献，引用格式为用方括号括起来的上标数字形式，按引用的次序依次顺序编号，并在报告末尾添加“参考文献”一节；每一条文献条目中至少应包括作者名，文章标题，期刊名，期号，卷号，出版年月，pp：页码范围，DOI 号或官网的 URL。

4. 必须提交的材料

- 下载的数据集：各个数据集各自存入一个文件中，文件名为程序中使用该数据集时的名称；
- python 的源程序：每个源程序存入一个文件，文件名能体现其作用；
- pdf 版本的技术报告；
- 以上三部分压缩成一个压缩包，以学号 + 姓名对压缩包进行命名。