同济大学软件学院 2021 年春季数据分析与数据挖掘课程作业 2

一．任务

1. 对指定的数据集进行聚类分析（链接 http://archive.ics.uci.edu/ml/datasets/Diabetes+130-US+hospitals+for+years+1999-2008）。(a) 数据之间的距离定义是分析过程中一个重要的部分，选择你认为合适的数据间距离定义；(b)在运行聚类算法的时候，需要设置一些参数，其中类的个数是重要的一个参数。对所选数据集进行分析来确定该数据集的类的个数；(c)从给定的聚类算法中任选两种进行实验比较分析（从效率和效果两方面），算法包括：K-Means, DBSCAN, Hierarchical, Spectral Clustering 和 EM-GMM 算法；(d) 选择合适的评价指标对不同算法的聚类结果进行评估， 并针对每一种算法记录最佳的聚类结果；(e) 综合以上几个方面，分析结果并写成报告。

2. 给定一份机顶盒数据集，其中，一个机顶盒卡号代表一户家庭，并将一户家庭作为一个用户。该数据集包括用户收藏记录和用户常看直播频道记录样例。两种数据集的具体格式如下：

1) 用户收藏记录

```
{
  ″CODE″：″媒资 ID″,
  ″FOLDERCODE″：″媒资所属栏目 ID″,
  ″NAME″：″媒资名称″,
  ″PORTAL_VER″：″互动版本″,
  ″SHOW_TYPE″：″媒资类型″,
  ″STBID″：″智能卡号″,
  ″TIME″：″收藏时间″
}
```

2) 用户常看直播频道记录样例

```
{
  ″SID″：″频道 serviceID″,
  ″OPK″：″区域码″,
  ″STBID″：″机顶盒卡号″,
  ″L_CHANNEL_NAME″：″频道名称″,
  ″CNT″：″观看次数″,
}
```

请以用户为主体对用户行为进行聚类分析，分析内容不做具体要求，自行发挥，并将分析结果和图表写成报告。

注：

1. **算法可以调用现有的实现，不会影响评分。**
2. **请保证报告内的分析具有一定的价值，不要做不必要的分析，避免报告过于冗长。**

二．提交

提交日期：2021-4-8 23:59，提交至 canvas。提交内容要求：

提交文件命名为学号_姓名(中文)_hw2.zip。共有两个子目录，对应两个任务，命名为 q1，q2，每个子目录包括以下内容：

1) 源代码文件。

2) README 文件，介绍运行环境和运行方式。

3) 实验报告文件，包括数据预处理、样本间距离定义，实验结果以及对实验结果的比较分析等。

4) 实验结果文件。任务 1 和 任务 2 各一个文件，均为 csv 文件格式。聚类结果用新的字段来表示。选择不同的聚类算法将对应不同的字段：

[kmeans_label, dbscan_label, hierarchical_label, spectral_label, em_label]. 每个字段的字段值用分类编号来表示，e.g. [0, 1, 1, 7, 12, 8]。