作业报告

学　号：201814841　　姓　名：徐　强

经过四周的学习和实践，基本完成了作业要求的内容。报告内容主要分为作业完成情况、学习收获和几点思考三个部分。

一、作业完成情况

按照聚类处理流程，将程序划分为四个步骤编写，具体实现过程如下：

（一）文档读取

Tweets 文件内容为 JSON 格式，打开文件后按行进行了读取，并将数据和聚类标签分别存放在列表结构中。

（二）计算 TF-IDF

与 KNN 中生成 VSM 方法类似，利用 Sklearn 库中的 CountVectorizer 和 TfidfTransformer 模块，计算出所有文本数据的的 TF 和 IDF，并生成 TF-IDF 权重矩阵。

（三）分别利用 7 种聚类方法进行聚类

1.K-means

1.1 算法

a.随机选择 k 个中心。

b.重复下列步骤直到达到停止条件停止条件：聚类中心不再发生变化或所有的距离最小或迭代次数达到设定值。

遍历所有样本，把样本划分到距离最近的一个中心。

计算每个簇的平均值作为新的质心。

1.2 算法评价

该算法聚类效果不错，也容易理解，速度快；但是需要自己确定 K 值，且初始中心的选取会影响最终的聚类结果。

2.Affinity Propagation

2.1 算法

AP 聚类是通过在样本对之间发送消息直到收敛来创建聚类。然后使用少量示例样本作为聚类中心来描述数据集，聚类中心是数据集中最能代表一类数据的样本。在样本对之间发送的消息表示一个样本作为另一个样本的示例样本的适合程度，适合程度值在根据通信的反馈不断更新。更新迭代直到收敛，完成聚类中心的选取，因此也给出了最终聚类。

2.2 算法评价

与 K-Means 等聚类算法不同的地方在于 AP 聚类不需要提前确定聚类的数量，即 K 值，但是运行效率较低。

3.Mean-shift 均值迁移

3.1 算法描述

Mean-shift 聚类的目的是找出最密集的区域，同样也是一个迭代过程。在聚类过程中， 首先算出初始中心点的偏移均值，将该点移动到此偏移均值，然后以此为新的起始点，继续移动，直到满足最终的条件。

3.2 算法评价

Mean-shift 也引入了核函数，用于改善聚类效果。除此之外，Mean-shift 在图像分割， 视频跟踪等领域也有较好的应用；但需要先用工具计算 bandwidth，如果设置不合适会影响聚类效果。

4.Spectral Clustering 谱聚类

4.1 算法描述

将样本看作顶点，样本间的相似度看作带权的边，从而将聚类问题转为图分割问题：找到一种图分割的方法使得连接不同组的边的权重尽可能低(这意味着组间相似度要尽可能低)，组内的边的权重尽可能高(这意味着组内相似度要尽可能高)。

4.2 算法评价

能够识别任意形状的样本空间且收敛于全局最优解。

5.Agglomerative Clustering 层次聚类

5.1 算法描述

自底向上的层次聚类。

初始时，所有点各自单独成为一类，然后采取某种度量方法将相近的类进行合并，并且度量方法有多种选择。合并的过程可以构成一个树结构，其根节点就是所有数据的集合，叶子节点就是各条单一数据。

AgglomerativeClustering 中可以通过参数 linkage 选择不

同的度量方法， 用来度量两个类之间的距离，可选参数有ward（类间距离等于两类对象之间的最小距离）、complete（类间距离等于两组对象之间的最大距离）、average（类间距离等于两组对象之间的平均距离）三个。

5.2 算法评价

可能会产生聚类结果得到的类的大小不均衡的结果。由于层次聚类涉及到循环计算，所以时间复杂度比较高，运行速度较慢。

6.DBSCN 密度聚类

6.1 算法

从某个选定的核心点出发，不断向密度可达的区域扩张，从而得到一个包含核心点和边界点的最大化区域，区域中任意两点密度相连。

6.2 算法评价

不需要指定 cluster 的数目，聚类的形状可以是任意的，能找出数据中的噪音，对噪音不敏感，聚类结果几乎不依赖于节点的遍历顺序；但需要设置合适的领域半径和最小核心点数量。

7.GaussianMixtureModel 混合高斯模型

7.1 算法描述

聚类算法大多数通过相似度来判断，而相似度又大多采用欧式距离长短作为衡量依据。而 GMM 采用了新的判断

依据：概率，即通过属于某一类的概率大小来判断最终的归属类别。

7.2 算法评价

GMM 的优点是投影后样本点不是得到一个确定的分类标记，而是得到每个类的概率，这是一个重要信息。GMM 不仅可以用在聚类上，也可以用在概率密度估计上。

（四）利用 NMI 方法进行评分

最终结果为：

```
The K-Means score is:0.769170
The Affinity Propogation score is:0.794534
The Mean-Shift score is:0.742985
The Spectral Clustering score is:0.781465
The Agglomerative Clustering-average score is:0.896244
The DBSCAN score is:0.654360
The Gaussian Mixtures score is:0.790958
```

二、学习收获

1.通过实验重新复习了老师课堂讲的 K-means 聚类、Agglomerative Clustering 层次聚类和 DBSCN 密度聚类三种方法，能够做到基本掌握，对 Mean-shift 均值迁移、Spectral Clustering 谱聚类、GMM 混合高斯模型、Affinity Propagation 和 BIRCH 五种方法有了初步的了解。

2.对 TF-IDF 的权重矩阵有了新的认识，可以用在对文本进行分类和聚类的多种任务中。

三、几点思考

1.层次聚类得分较高，密度聚类得分较低，有可能和密

度聚类的领域半径和最小核心点数量的设置有关。

2.从程序运行时间来看，谱聚类最快，GMM 时间最长，GMM 方法收敛速度慢主要原因可能为数据不足时估算协方差矩阵困难，同时算法会发散并且找具有无穷大似然函数值的解，需人为地对协方差进行正则化。