Clustering with Sklearn

一、本次实验主要是验证 sklearn 中的几种聚类算法在 tweets 数据集上的聚类效果。

二、Project flow

三、几种不同聚类方法的比较

1. KMeans

1.1 算法描述

1. 随机选择 k 个中心
2. 遍历所有样本，把样本划分到距离最近的一个中心
3. 划分之后就有 K 个簇，计算每个簇的平均值作为新的质心
4. 重复步骤 2，直到达到停止条件

停止条件：

聚类中心不再发生变化；所有的距离最小；迭代次数达到设定值

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。

2. DBSCN

2.1 算法描述

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。

2.2 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; 但是如果样本集较大时, 聚类收敛时间较长，较耗计算资源。

3. Spectral Clustering 谱聚类

3.1 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)。

3.2 算法评价

能够识别任意形状的样本空间且收敛于全局最有解

4. Agglomerative Clustering 层次聚类

4.1 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。

sklearn. cluster. 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, complete, average 三个。

Ward（WardHierarchicalClustering）：选择这样的两个类进行合并，合并后的类的离差平方和最小。

4.2 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。

5. Mean-shift 均值迁移

5.1 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。

5.2 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。

6. GaussianMixtureModel (混合高斯模型)

6.1 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。

6.2 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。

但是当每个混合模型没有足够多的点时，估算协方差变得困难起来，同时算法会发散并且找具有无穷大似然函数值的解，除非人为地对协方差进行正则化。GMM 每一步迭代的计算量比较大，大于 k-means。

7. Affinity Propagation

7.1 算法描述

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。

7.2 算法评价

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

四、实验结果及问题

1. 实验结果

截图：

```
==============KMeans==============
NMI: 0.7821292248361797
=======AffinityPropagation========
NMI: 0.7836988975391974
===========MeanShift=============
NMI: -1.6132928326584306e-06
====WardHierarchicalClustering====
NMI: 0.7847178748775534
=======SpectralClustering=========
NMI: 0.6619084064007756
=====AgglomerativeClustering=====
NMI: 0.7847178748775534
============DBSCAN===============
NMI: 0.611995415736244
=========GaussianMixture==========
NMI: 0.7798759561032018
```

MeanShift 的速度最慢，从 NMI 评价来看 AgglomerativeClustering 的聚类效果最好。

2. 问题

1）Tweets 文件为 json 格式，在读取文件的时候有一些问题。一开始读取的时候，不知道可以直接调用读取 json 文件的方法，就采用了手动分开 text 和 cluster 的方法，并将其分别存在两个 list 中。

2）GMM 方法没有 fit_predict 方法，与之对应的是 predict 方法。

3）fit 方法与 fit_predict 方法探究：fit 方法返回的是一个聚类模型，而 fit_predict 返回的直接是聚类的 label 结果。