山东大学 计算机科学与技术 学院

信息检索 课程实验报告

学号：201600150109	姓名：沈棋韬	班级： 16 人工智能班

实验题目：Clustering with sklearn

实验内容：

测试 sklearn 中以下聚类算法在 tweets 数据集上的聚类效果。

使用 NMI（Normalized Mutual Information）作为评价指标。

实验过程中遇到和解决的问题：

（记录实验过程中遇到的问题，以及解决过程和实验结果。可以适当配以关键代码辅助说明，但不要大段贴代码。）

在聚类前，需要对数据集进行预处理。首先需要把 tweets 从文档中读出，并保存在字典中。然后建立一个标签数组，来保存每一条 tweet 的 cluster 标签，再建立一个列表，保存每一条 tweet 中的每一个单词。因为 sklearn 需要输入的数据集为 np 数组，所以再对每条 tweet 进行向量化。把所有出现过的词从 0 开始编号，每一条推特中若包含这个单词则该位置为 1，否则为 0。

这样以后就可以开始聚类。

Kmeans：

```
In [7]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster_number:  110
Kmeans socre:  0.7413874624948446
```

因为 tweets 中有 110 个类，所以 Kmeans 的 cluster 数量设置为 110。再使用 NMI 评价，结果如图所示。NMI 的结果越接近 1，表示聚类的结果越好。

Affinity propagation：

```
In [8]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster_number:  163
affinity propagation socre:  0.7468614810451298
```

使用默认参数，数据集被分成了 163 个 cluster。

Mean-shift：

```
In [7]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster_number:  6
meanShift score:  0.03106110377739007
```

Mean-shift 中，使用默认参数，划分成了 6 个 cluster，耗费时间较长，NMI 得分较低。

spectral clustering:

```
In [14]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluter number:  8
spectral clustering socre:  0.08374142407520367
```

使用默认参数，分为 8 个 cluster。

Ward hierarchical clustering:

```
In [16]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  110
Ward hierarchical clustering socre:  0.8188750199403455
```

人为给参数 110，得到结果较好。

Agglomerative clustering:

```
cluster number:  110
Agglomerative average clustering score:  0.18583004008382129
```

使用参数 linkage=average，给定 cluster 数量 110，NMI 分数并不好。

```
In [20]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  110
Agglomerative complete clustering score:  0.7280657189516214
```

使用参数 linkage=complete，给定 cluster 数量 110，NMI 得分较好。

DBSCAN:

```
In [23]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  5
DBSCAN score:  0.0914960470508469
```

使用默认参数，划分成了 5 个 cluster，NMI 得分不高。

Gaussian mixtures：

```
In [1]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  2
Gaussian mixtures score:  0.09909604119040356

In [2]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  3
Gaussian mixtures score:  0.18625658329128064

In [3]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  4
Gaussian mixtures score:  0.35552702092321453

In [4]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  5
Gaussian mixtures score:  0.3682052752283387

In [5]: runfile('D:/pythonCode/实验5聚类.py', wdir='D:/pythonCode')
cluster number:  6
Gaussian mixtures score:  0.2861113630109515

In [6]:
```

在高斯混合模型中，模型数量默认为 1，通过人为改变参数，每次的结果不同，且使用模型数量越多，计算越慢。在第五次时达到了最高的 NMI 得分。

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n_samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n_samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n_samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n_samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers

通过上图，以及实验的结果。可知 K-means， affinity propagation， Ward hierarchical clustering， Agglomerative clustering 对于较大的数据集，在 cluster 数较多的时候效果较好。而 Mean-shift， spectral clustering, DBSCAN 和 Gaussian mixture 不适合在较多 cluster 时使用。