UCLA CS 145 Homework #3

DUE DATE: Friday 11/10/2017 11:59 pm

Note:

- You are expected to submit both report and code. For your code, please include clear readme files. When not specifically mentioned, use the default data provided in our provided program package.
- "`# **********Please Fill Missing Lines Here**********`" is used where input from you is needed.

1. Clustering Evaluation.

ID	Conference Name	Ground Truth Label	Algorithm output Label
1	IJCAI	3	2
2	AAAI	3	2
3	ICDE	1	3
4	VLDB	1	3
5	SIGMOD	1	3
6	SIGIR	4	4
7	ICML	3	2
8	NIPS	3	2
9	CIKM	4	3
10	KDD	2	1
11	WWW	4	4
12	PAKDD	2	1
13	PODS	1	3
14	ICDM	2	1
15	ECML	3	2
16	PKDD	2	1
17	EDBT	1	2
18	SDM	2	1
19	ECIR	4	4
20	WSDM	4	4

Suppose we want to cluster 20 above conferences into four areas, with ground truth label and algorithm output label shown in third and fourth column. Please evaluate the quality of the clustering algorithm according to purity, precision, recall, F-measure, and normalized mutual information, respectively.

2. K-means

(1) Fill in the missing lines in KMeans.py and run the algorithm against three datasets (dataset1.txt, dataset2.txt, and dataset3.txt), respectively. Please view the file README.txt for coding requirements.

(2) Plot the clustering results for the three datasets using a scatter plot, with different colors representing different clusters. Evaluate the algorithm using (1) purity and (2) normalized mutual information for each dataset.

(3) Give the strengths and weaknesses of using the K-means algorithm.

3. DBSCAN

(1) Fill in the missing lines in DBSCAN.py and run the algorithm against three datasets (dataset1.txt, dataset2.txt, and dataset3.txt), respectively. Please view the file README.txt for coding requirements.

(2) Plot the clustering results for the three datasets using a scatter plot, with different colors representing different clusters. Evaluate the algorithm using (1) purity and (2) normalized mutual information for each dataset.

(3) Give the strengths and weaknesses of using DBSCAN.

3. GMM

(1) Fill in the missing lines in GMM.py and run the algorithm against three datasets (dataset1.txt, dataset2.txt, and dataset3.txt), respectively. Please view the file README.txt for coding requirements.

(2) Plot the clustering results for the three datasets using a scatter plot, with different colors representing different clusters. Evaluate the algorithm using (1) purity and (2) normalized mutual information for each dataset.

(3) Give the strengths and weaknesses of using GMMs.