CS6220 Data Mining Fall 2014 Homework 2, Wei Luo

1. Clustering Evaluation

The clusters and ground truth label can be summarized as:

cluster	1	2	3	4
labels	{2,2,2,2,2,4}	{3,3,3,3,3,1}	{1,1,1,4,1}	{4,4,4}

$purity = \frac{1}{20} * (5+5+4+3) = 0.85$
$TP = \binom{5}{2} + \binom{5}{2} + \binom{4}{2} + \binom{3}{2} = 29$
$FP = 5+5+4 = 14$
$FN = 4+0+0+7 = 11$
$TN = 5*14 + 1*10 + 5*8 + 1*4 + 4*3 = 136$
precision: $P = \frac{TP}{TP+FP} = \frac{29}{29+14} = 0.6744$
recall: $R = \frac{TP}{TP+FN} = \frac{29}{29+11} = 0.725$
F-measure: $F_1 = \frac{2PR}{P+R} = \frac{2*0.6744*0.725}{0.6744+0.725} = 0.6988$
For clusters C and ground true labels Ω:

$$I(\Omega, C) = \frac{1}{20}\log\frac{20*1}{5*6} + \frac{4}{20}\log\frac{20*4}{5*5} + \frac{5}{20}\log\frac{20*5}{5*6} + \frac{5}{20}\log\frac{20*5}{5*6}$$
$$+\frac{1}{20}\log\frac{20*1}{5*6} + \frac{1}{20}\log\frac{20*1}{5*5} + \frac{3}{20}\log\frac{20*3}{5*3} = 0.9909$$
$$H(\Omega) = -(\frac{5}{20}\log\frac{5}{20} + \frac{5}{20}\log\frac{5}{20} + \frac{5}{20}\log\frac{5}{20} + \frac{5}{20}\log\frac{5}{20}) = 1.3863$$
$$H(C) = -(\frac{6}{20}\log\frac{6}{20} + \frac{6}{20}\log\frac{6}{20} + \frac{5}{20}\log\frac{5}{20} + \frac{3}{20}\log\frac{3}{20}) = 1.3535$$

Normalized Mutual Information: $NMI(\Omega, C) = \frac{I(\Omega,C)}{\sqrt{H(\Omega)H(C)}} = 0.7234$

2. Understanding and comparing different clustering algorithms.

(1) run the code with commands like:

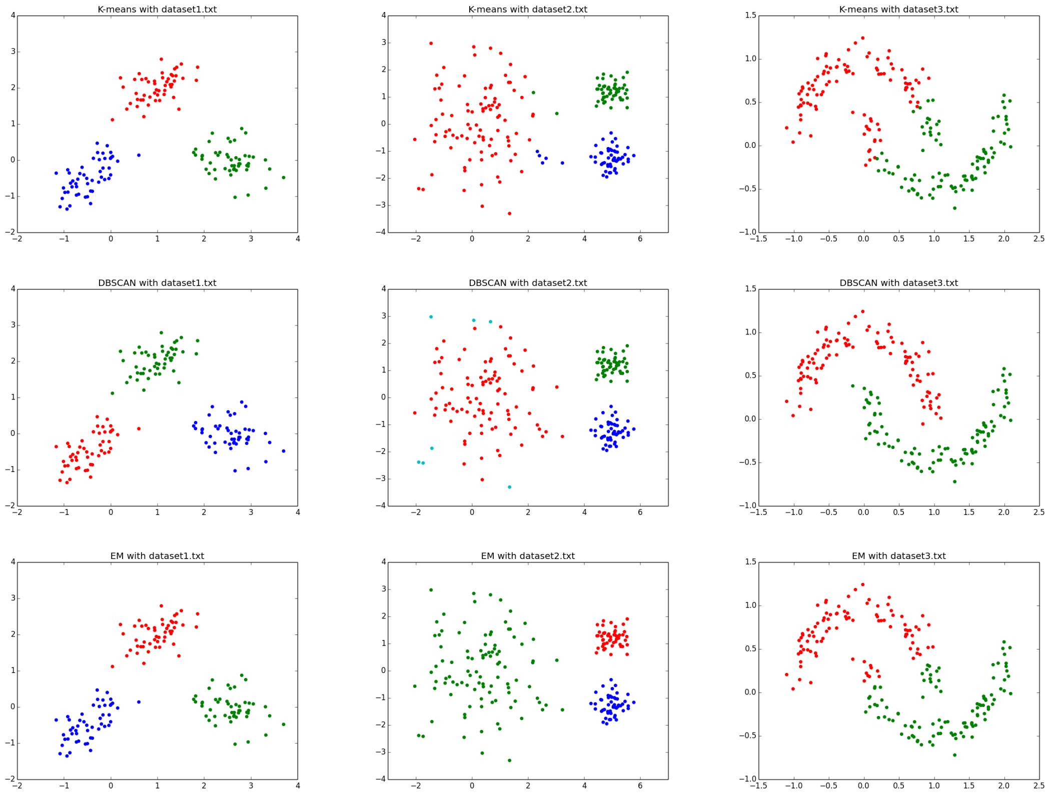
```
> python k_means.py dataset1.txt
> python em.py dataset2.txt
> python dbscan.py dataset3.txt 0.3 10
```

Where the first parameter (dataset*.txt) is the file name of the dataset. For DBSCAN, the second parameter is *eps*, the third parameter is *minPts*
(2) The purity and NMI of each algorithm for each dataset is:

		dataset1.txt	dataset2.txt	dataset3.txt
K-means	Purity	1.000000	0.965000	0.820000
	NMI	1.000000	0.869672	0.319923
DBSCAN	Purity	1.000000	0.965000	1.000000
	NMI	1.000000	0.922051	1.000000
EM	Purity	1.000000	1.000000	0.880000
	NMI	1.000000	1.000000	0.472501

In DBSCAN algorithm, I used
$eps = 0.7, minPts = 15$ for dataset1
$eps = 0.9, minPts = 5$ for dataset2
$eps = 0.3, minPts = 10$ for dataset3

The cluster result as scatter plot is:



(3) For dataset1, all three algorithm successfully clustered the clusters with 100% accuracy. Because the datapoints are very well separated, all algorithms can do the job well.

For dataset2, the EM algorithm works better, because although the clusters are linearly separable, they have different shapes. The cluster on the left has much bigger range area and data points are sparse distributed. This will prevent K-means algorithm working well where each cluster should have almost same size. Also, this will introduce noise in DBSCAN algorithm. But for EM with gaussian mixtures, the gaussians can have different shape which can make a perfect fit to dataset2 clusters.

For dataset3, the clusters are not linearly separable, which makes the result of K-means and EM algorithms hard to do the clustering. But the data points are densely distributed. This makes DBSCAN algorithm works better than the other two algorithms.

3. Clustering the real-world data.

(1) Major steps:

1. By looking into the data, find out the top 10% authors who have most contributions in the 20 conferences. (Have most publications in those conferences.)

2. Feature selection: take the count of each top 10% authors' publications in each conference as features. The number of features is the number of authors who have top 10% contributions among conferences.

3. Clustering criterion: for the N features, we look into the N dimensional space, among the 20 data points, the ones that are closer to each other should be considered as in same cluster.

4. Algorithm to use: choose the K-means algorithm with $k = 4$ to do the clustering.

(2) Purity is 0.750000; NMI is 0.340576.

(3) By assuming that the authors are more likely to focus on similar topics and publish their publications to similar conferences, this method is reasonable. However, looking at the result, this method can only get a fair purity and low NMI. This indicates that the previous assumption may not stand.