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

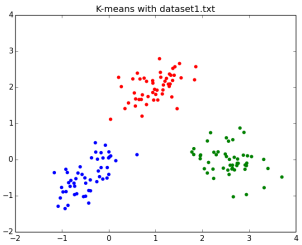
2. Understanding and comparing different clustering algorithms.

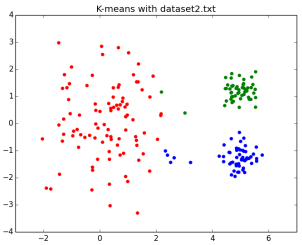
(1) run the code with commands like:
> python k_means.py dataset1.txt
> python emp.y dataset2.txt
> python dbscan.py dataset3.txt 0.3 10
Where the first parameter (dataset*.txt) is the file name of the dataset. For DBSCAN, the second parameter is *eps*, the third parameter is $minPts$
(2) The purity and NMI of each algorithm for each dataset is:

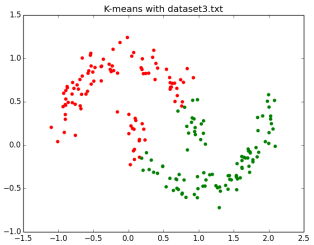
		dataset1.txt	dataset2.txt	dataset3.txt
K-means	Purity	1.000000	0.965000	0.820000
	NMI	1.000000	0.869672	0.319923
DBSCAN	Purity	1.000000	0.965000	1.000000
	NMI	1.000000	0.922051	1.000000
EM	Purity	1.000000	1.000000	0.880000
	NMI	1.000000	1.000000	0.472501

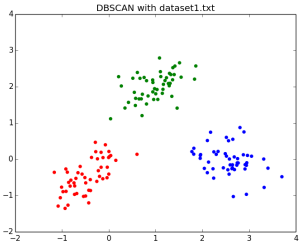
In DBSCAN algorithm, I used
$eps = 0.7, minPts = 15$ for dataset1
$eps = 0.9, minPts = 5$ for dataset2
$eps = 0.3, minPts = 10$ for dataset3

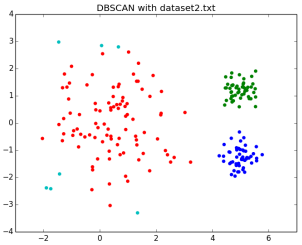
The cluster result as scatter plot is:

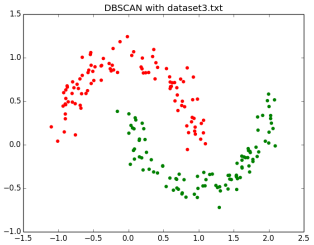

K-means with dataset1.txt

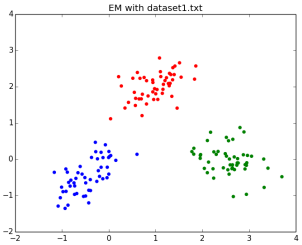

K-means with dataset2.txt

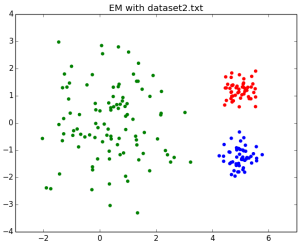

K-means with dataset3.txt

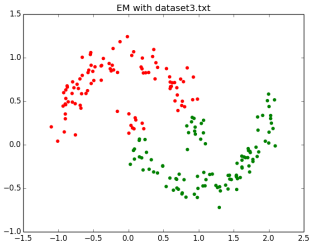

DBSCAN with dataset1.txt


DBSCAN with dataset2.txt


DBSCAN with dataset3.txt


EM with dataset1.txt


EM with dataset2.txt


EM with dataset3.txt