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:
$I(\Omega, C) = \frac{1}{20}\log\frac{20*1}{5*6} + \frac{4}{20}\log\frac{20*4}{5*5} + \frac{5}{20}\log\frac{20*5}{5*6} + \frac{5}{20}\log\frac{20*5}{5*6}$
$+\frac{1}{20}\log\frac{20*1}{5*6} + \frac{1}{20}\log\frac{20*1}{5*5} + \frac{3}{20}\log\frac{20*3}{5*3} = 4.2753$
$H(\Omega) = -(\frac{5}{20}\log\frac{5}{20} + \frac{5}{20}\log\frac{5}{20} + \frac{5}{20}\log\frac{5}{20} + \frac{5}{20}\log\frac{5}{20}) = 1.3863$
$H(C) = -(\frac{6}{20}\log\frac{6}{20} + \frac{6}{20}\log\frac{6}{20} + \frac{5}{20}\log\frac{5}{20} + \frac{3}{20}\log\frac{3}{20}) = 1.3535$
Normalized Mutual Information: $NMI(\Omega, C) = \frac{I(\Omega,C)}{\sqrt{H(\Omega)H(C)}} = 3.1211$