Worksheet 09

Name: Ian Tsai, Sangheon Jeong UID: U10536401, U72771619

Topics

- Clustering Review
- Clustering Aggregation

Clustering Aggregation

Point	C	P
A	0	a
B	0	b
C	2	b
D	1	c
E	1	d

a) Fill in the following table where for each pair of points determine whether C and P agree or disagree on how to cluster that pair.

Pair	Disagreement
A B	1
A C	0
A D	0
A E	0
B C	1
B D	0
B E	0
C D	0

Pair	Disagreement
C E	0
D E	1

As datasets become very large, this process can become computationally challenging.

b) Given N points, what is the formula for the number of unique pairs of points one can create?

The formula is "n(n-1)/2"

Assume that clustering C clusters all points in the same cluster and clustering P clusters points as such:

Point	P
A	0
B	0
C	0
D	1
E	1
F	2
G	2
H	2
I	2

c) What is the maximum number of disagreements there could be for a dataset of this size? (use the formula from b)?

(9*(9-1))/2 = 36

d) If we look at cluster 0. There are (3 x 2) / 2 = 3 pairs that agree with C (since all points in C are in the same cluster). For each cluster, determine how many agreements there are. How many total agreements are there? How many disagreements does that mean there are between C and P?

For cluster 0, there are 3 agreements (A, B, C). For cluster 1, (2x1)/2=1. For cluster 2, (4x3)/2=6. Hence the total agreement is 10. For total disagreement, we do 36-10 = 26 disagreements.

e) Assuming that filtering the dataset by cluster number is a computationally easy operation, describe an algorithm inspired by the above process that can efficiently compute disagreement distances on large datasets.

Aggregate clustering, k-means

In [ ]: