worksheet_09

April 30, 2024

1 Worksheet 09

Name: Afiq Amjad bin Khairir UID: U41760804

1.0.1 Topics

- Clustering Review
- Clustering Aggregation

1.0.2 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	Disagree
A C	Agree
A D	Agree
A E	Agree
B C	Disagree
B D	Agree
B E	Agree
C D	Agree
C E	Agree
D E	Disagree

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?

$$\text{Number of unique pairs} = \binom{N}{2} = \frac{N(N-1)}{2}$$

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 * 8 / 2 = 72 / 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 1: (2 * 1) / 2 = 1 For cluster 2: (4 * 3) / 2 = 6

Total agreement: 10 Maximum disagreements: 36

Disagreements between C and P: 36 - 10 = 26

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.

1. Filter the dataset by cluster number.
2. Find intersection of data points in C and P that are in the same cluster.
3. For each grouping, use the N(N-1)/2 formula and total it together to get the total agreements.
4. Calculate the total number of disagreements using the same formula.
5. Subtract total number of disagreements with the total number of agreements to get the disagreement distance.