Clustering Aggregation

Clustering Aggregation

Some terminology:

Clustering: A group of clusters output by a clustering algorithm

Cluster: A group of points

Clustering Aggregation

Goals:

1. Compare clusterings
2. Combine the information from multiple clusterings to create a new clustering

Comparing Clusterings

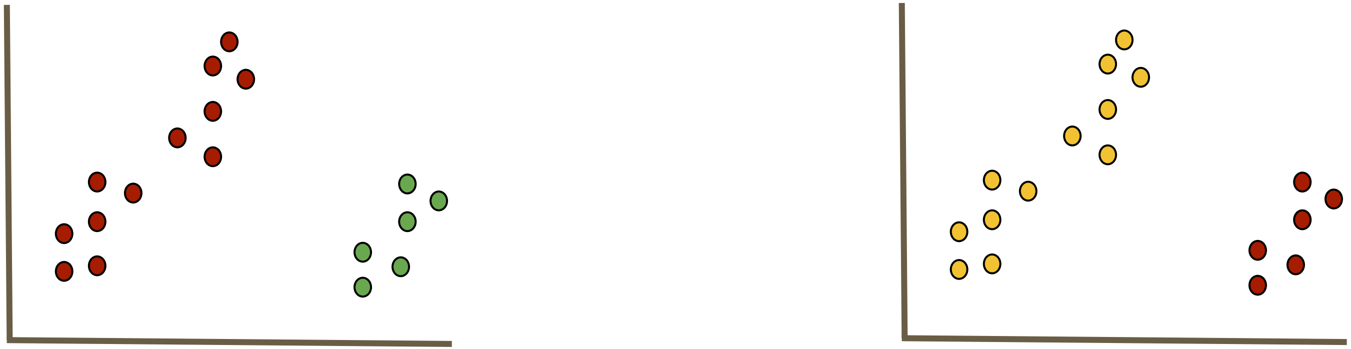
The many methods / cost functions make comparing clusterings difficult.

Need to compare clusterings by looking at their assignment of points to clusters.

If many points were assigned to the same clusters in both clustering C and clustering P, then C and P should have a small distance.

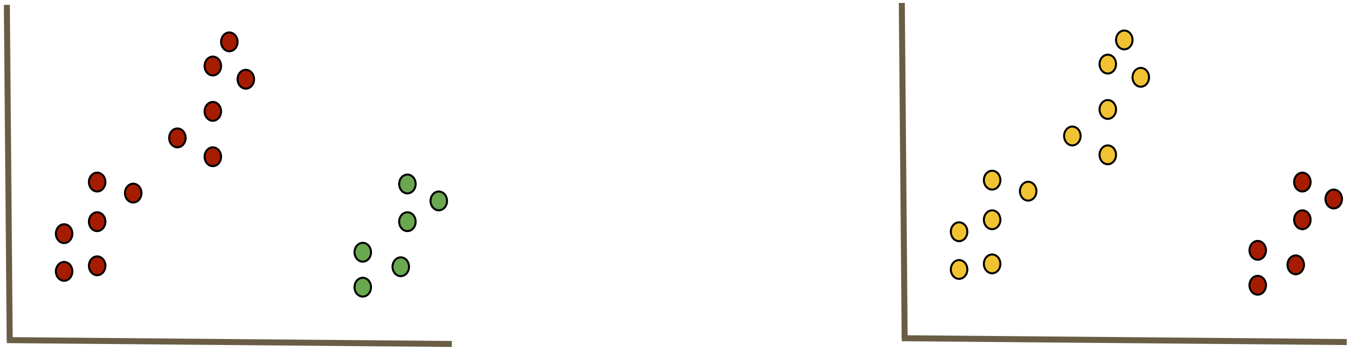
But identifying which clusters are the same in P and C is not easy. Why?

Comparing Clusterings



Clearly these clusterings are the same. Yet the assignments / labels are inconsistent.

Comparing Clusterings



Asking “is x in cluster “red”” in the left clustering is equivalent to asking “is x in cluster “yellow”” on the right clustering but we cannot know this conversion up front unless there is a known set of conventions.

Comparing Clusterings

Let’s not limit ourselves with such a set of convention and instead ask a different question:

Are x and y clustered together in both P and C?

Disagreement Distance

Given 2 clusterings P and C

$$D(P,C) = \sum_{x,y} \mathbb{I}_{P,C}(x,y)$$

where

$$\mathbb{I}_{P,C}(x,y) = \begin{cases} 1 & \text{if P \& C disagree on which clusters x \& y belong to} \\ 0 & \end{cases}$$

Disagreement Distance

	P	C
x_1	1	1
x_2	1	2
x_3	2	1
x_4	3	3
x_5	3	4

What is the disagreement distance between P and C?

Disagreement Distance

	P	C
x_1	1	1
x_2	1	2
x_3	2	1
x_4	3	3
x_5	3	4

x_2	x_1	1
x_3	x_1	1
x_4	x_1	0
x_5	x_1	0
x_3	x_2	0
x_4	x_2	0
x_5	x_2	0
x_4	x_3	0
x_5	x_3	0
x_4	x_5	1

Disagreement Distance

Is D(P, C) a distance function?

1. D(C, P) = 0 iff C = P
2. D(C, P) = D(P, C)
3. Triangle Inequality:

$$\mathbb{I}_{C_1,C_3}(x,y) \leq \mathbb{I}_{C_1,C_2}(x,y) + \mathbb{I}_{C_2,C_3}(x,y)$$

Since $\mathbf{I}_{\mathbf{C,P}}$ can only be 0 or 1, the above can only be violated if

$\mathbf{I}_{\mathbf{x,y}}\mathbf{(C_1,C_3) = 1}$ **,** $\mathbf{I}_{\mathbf{x,y}}\mathbf{(C_1,C_2) = 0}$ **,** $\mathbf{I}_{\mathbf{x,y}}\mathbf{(C_2,C_3) = 0}$ is this possible?

Aggregate Clustering

Goal: From a set of clusterings $\mathbf{C_1}$**, …,** $\mathbf{C_m}$, generate a clustering $\mathbf{C^*}$ that minimizes:

$$\sum_{i=1}^{m} D(C^*, C_i)$$

The problem is equivalent to clustering categorical data

Aggregate Clustering

	City	**Profession**	**Nationality**
$\mathbf{x}_1$	NY	Doctor	US
$\mathbf{x}_2$	NY	Teacher	French
$\mathbf{x}_3$	Boston	Lawyer	Canada
$\mathbf{x}_4$	Boston	Doctor	US
$\mathbf{x}_5$	LA	Lawyer	Canda
$\mathbf{x}_6$	LA	Actor	French

Aggregate Clustering

Benefits:

1. Can identify the best number of clusters (optimization function does not make any assumptions on the number of clusters)
2. Can handle / detect outliers (points where there is no consensus)
3. Improve robustness of the clustering algorithms - combining clusterings can produce a better result
4. Privacy preserving clustering (can compute aggregate clustering without sharing the data, need only share the assignments)

Aggregate Clustering

But... The problem is NP-Hard.

Often use approximations and heuristics to solve this problem.

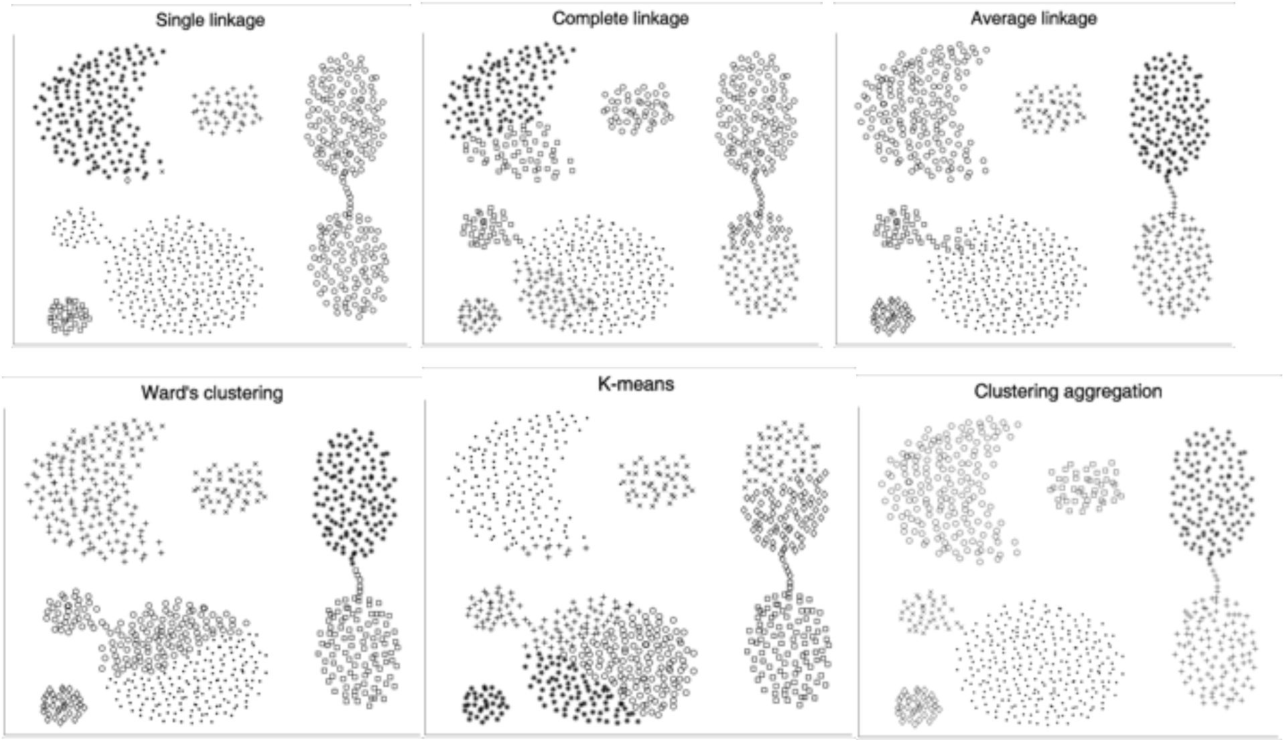
What about the majority rule?

This only works **if** it produces a clustering

Possible to have a majority saying:

1. x_1 & x_2 together
2. x_2 & x_3 together
3. x_1 & x_3 separate

Aggregate Clustering


Single linkage
Complete linkage
Average linkage
Ward's clustering
K-means
Clustering aggregation