Lecture 8: Clustering Aggregation

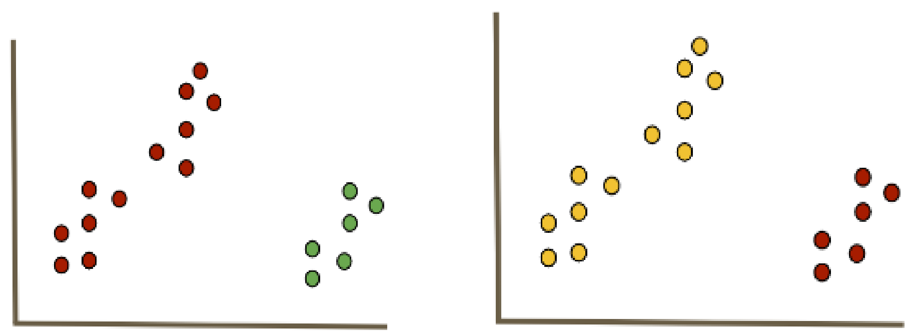
1) Clustering Aggregation

a) Terminology:
   i) Clustering: group of clusters output by clustering algorithm
   ii) Cluster: group of points

b) Goals:
   i) Compare clusterings
   ii) Combine information from multiple clusterings to create a new clustering

2) Comparing Clusterings

a) Need to compare clustering by looking at assignment of points in clusters
   i) Many points assigned to same cluster in both clustering C and P, then they should have a small distance
   ii) Identifying which cluster in P and C are not easy

b) 
   i) Clusterings are the same, but assignments/labels not consistent
   ii) Asking "is x in red cluster" in left clustering = "is x in yellow cluster" in right clustering
      1) However, won't know conversion unless we know set of conventions

3) Disagreement Distance

a) Given 2 clusterings P and C

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

i) Where

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

b) Ex:

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

i) Disagreement distance for P and C

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

c)

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

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

i) $I_{C,P}$ can only be 0 or 1 and the above is violated iff

$I_{x,y}(C_1,C_3) = 1$, $I_{x,y}(C_1,C_2) = 0$, $I_{x,y}(C_2,C_3) = 0$

4) Aggregate Clustering

a) Goal: From set of clusterings $C_1, \ldots, C_m$ generate a clustering C^* that minimizes

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

   i) Problem equivalent to clustering categorical data

b) Benefits:

   i) Identify best number of clusters

      1) Optimization function not make assumptions on number of clusters

   ii) Handle/detect outliers

   iii) Improve robustness of clustering algorithms -> combining clusters can produce better results

   iv) Privacy preserving clustering: aggregate clustering without sharing data

c) NP-Hard problem

   i) Often solve with approximations

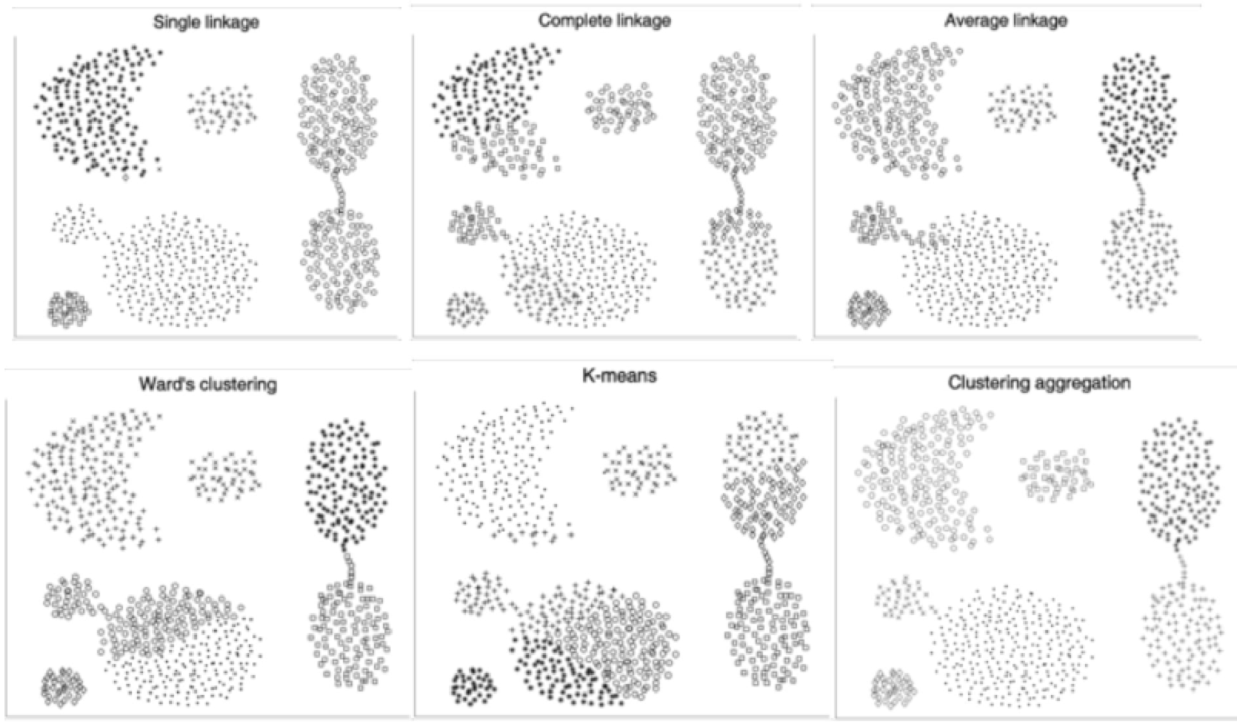
   ii) Rule: only worlds if it produces clustering

d)

	City	Profession	Nationality
x_1	NY	Doctor	US
x_2	NY	Teacher	French
x_3	Boston	Lawyer	Canada
x_4	Boston	Doctor	US
x_5	LA	Lawyer	Canda
x_6	LA	Actor	French

   i) Majority saying

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


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