CIS501 – Lecture 13

Woon Wei Lee
Fall 2013, 10:00am-11:15am,
Sundays and Wednesdays

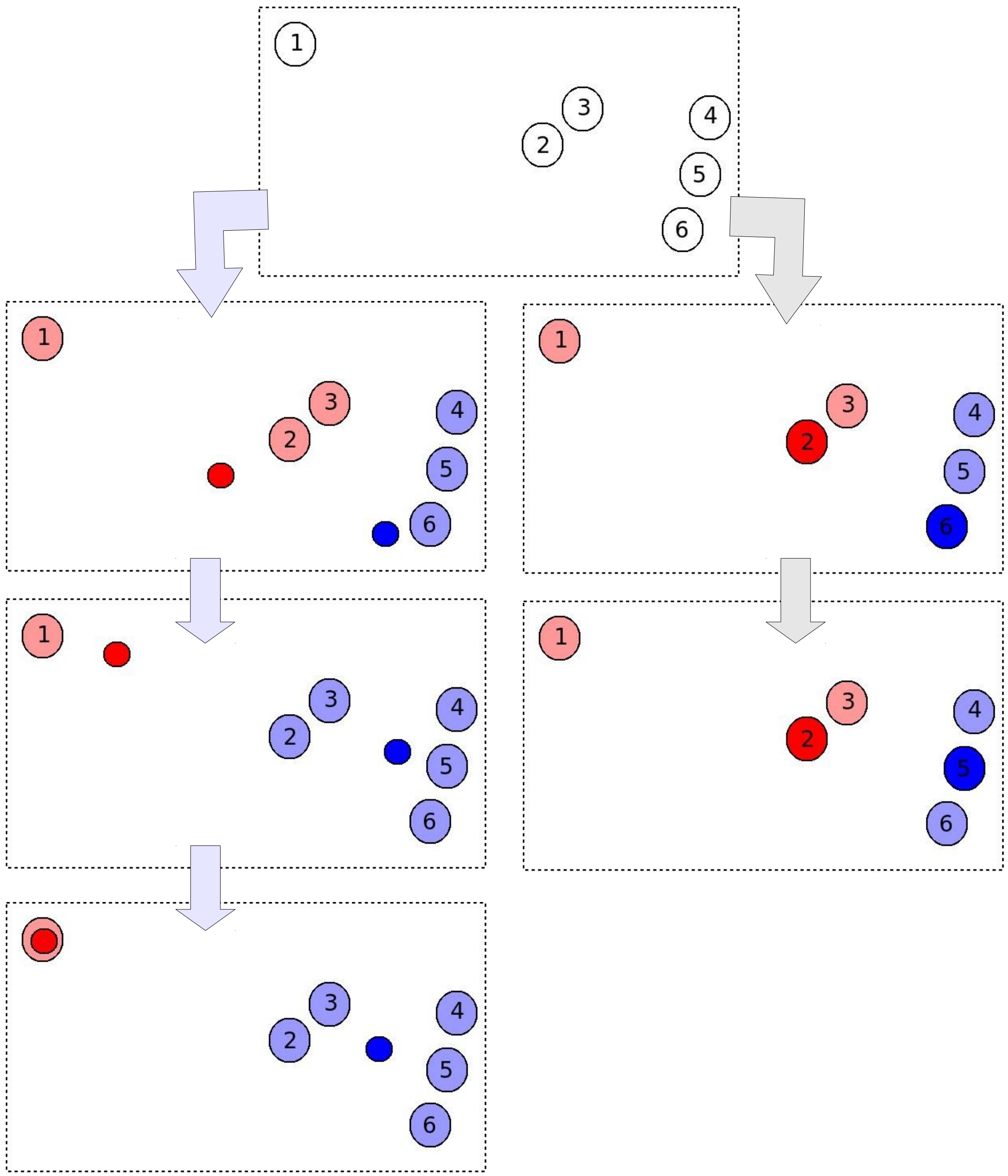
Masdar INSTITUTE

For today:

- Unsupervised learning
  - Cluster analysis
  - Quality measures
- Presentations
  - Artur Grigoryan
  - Chih-Hsien Chou

  (I think..)

K-means ↔ K-centers comparison



- In k-centers, you are using the *medoid* instead of the *mean* vector.
- Similar to mean/median division

  → outlier resistance

Clustering quality measures

- **Why?**
  - A means of validation – does clustering work at all?
    - Difficult to tell with high dimensional data!
  - Model order selection..
  - Clustering algorithms are often stochastic – can repeat and choose best outcome
  - Allows direct optimization of cluster partitions
- **Dunn index**

$$DI(c) = \min_{i,j \in c : i \neq j} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c} \left(\Delta(A_k) \right)} \right\}$$

(Large *DI* is good!)

  - $\delta(A_i, A_j)$ is the distance between the two closest points in clusters *i* and *j*
  - $\triangle(A_i)$ is the cluster “diameter”: i.e. the distance between the two furthest points in cluster *i*.

Quality measures (cont'd)

- **Davies-Bouldin Index**

$$DB(c) = \frac{1}{|c|} \sum_{i \in c} \max_{i \neq j} \left\{ \frac{\Delta(A_i) + \Delta(A_j)}{\delta(A_i, A_j)} \right\}$$ (Small *DB* is good!)

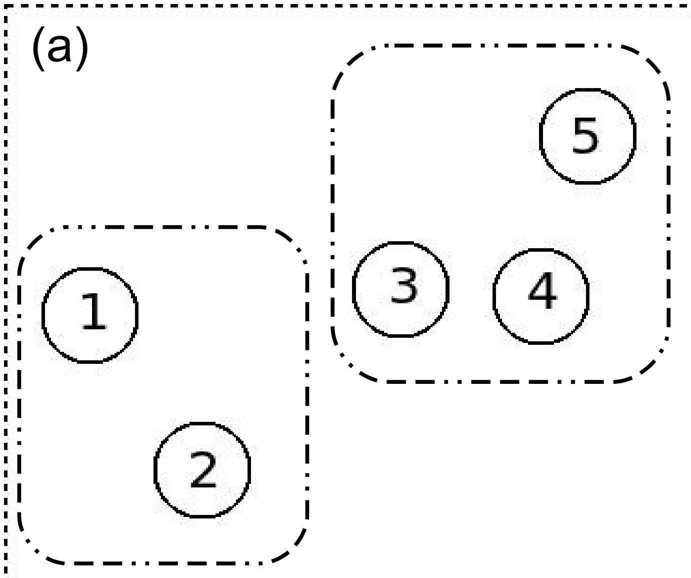
  - $\triangle(A_i)$ and $\delta(A_i,A_j)$ have the same meanings as in previous formula

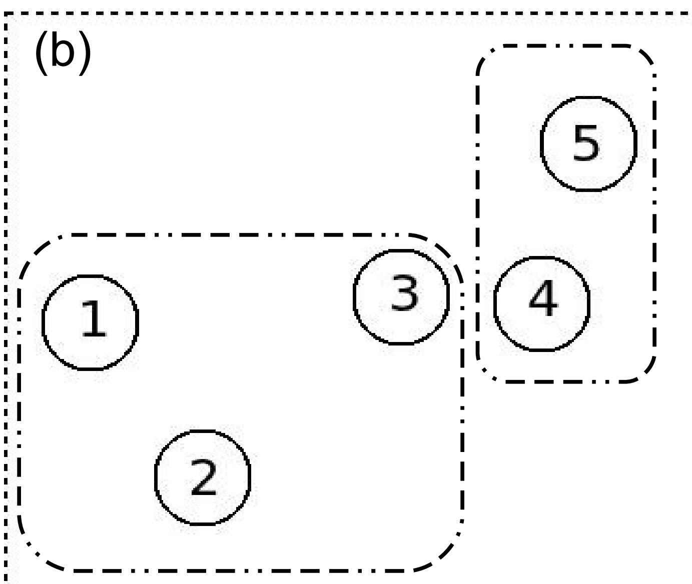
- **C-index**

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}$$ (Small C is good!)

  - *S* – sum of distances between all pairs of objects which are in the same cluster(s)
  - S_{min} – sum of the *n* smallest distances between all pairs of objects
  - S_{max} – sum of the *n* biggest distances between all pairs of objects

Example





	1	2	3	4	5
1	0	1	3	4	5
2	1	0	2	3	4
3	3	2	0	1	2
4	4	3	1	0	1
5	5	4	2	1	0

Dunn index:

(a) $$DI(c)_{(a)} = \min_{i,\, j \in c : i \neq j} \left\{ \frac{\delta(A_i, A_j)}{\max_{k \in c} \left(\Delta(A_k) \right)} \right\}$$

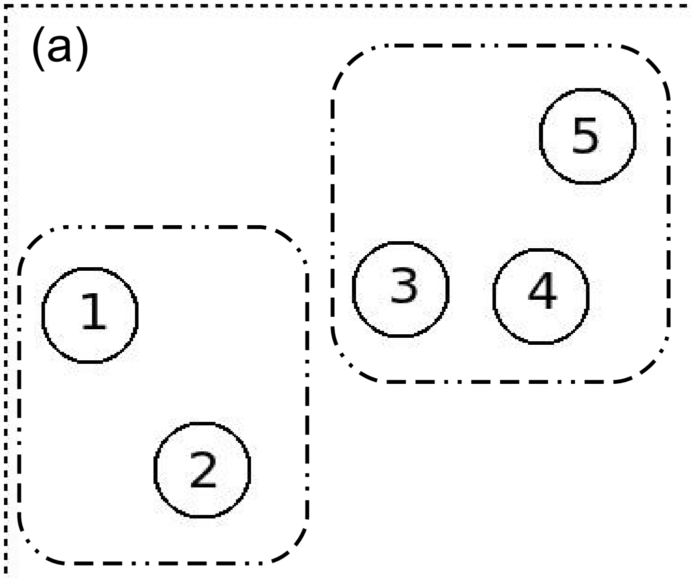
$$= \frac{2}{max\{1,2\}} = 1$$

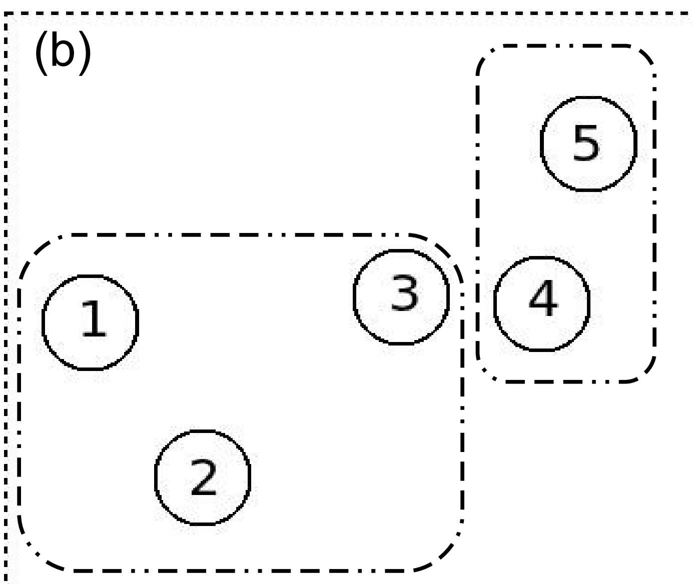
(b) $$DI(c)_{(b)} = \frac{1}{max\{3,1\}}$$

$$= \frac{1}{3} < 1$$

i.e. case (a) is the “better” clustering

Cont'd





Davies-Bouldin index

(a)

$$DB(c)_{(a)} = \frac{1}{|c|} \sum_{i \in c} \max_{i \neq j} \left\{ \frac{\Delta(A_i) + \Delta(A_j)}{\delta(A_i, A_j)} \right\}$$

$$= \frac{1}{2}\left[\frac{1+2}{2}\right] = 0.75$$

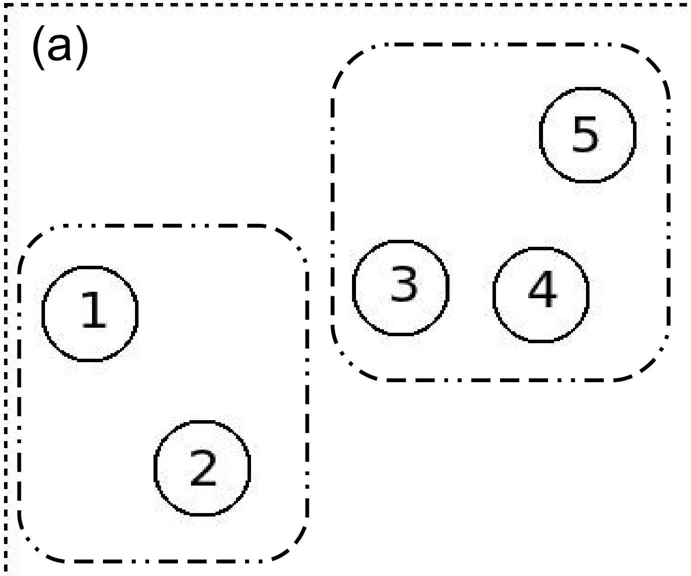
(b)

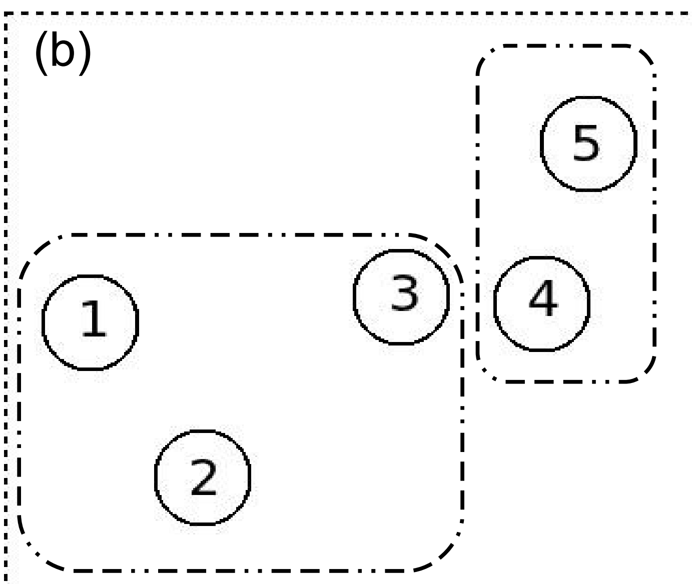
$$DB(c)_{(b)} = \frac{1}{2}\left[\frac{3+1}{1}\right]$$

$$= 2 > 0.75$$

(DB index → smaller the better)

C-index





(a)

$$C = \frac{S - S_{min}}{S_{max} - S_{min}}$$

$$S = 1 + (1 + 1 + 2) = 5$$

$$S_{min} = d_{12} + d_{34} + d_{45} + d_{35}$$
$$= 1 + 1 + 1 + 2 = 5$$

$$S_{max} = d_{15} + d_{25} + d_{14} + d_{24}$$
$$= 5 + 4 + 4 + 3 = 16$$

$$C_{(a)} = \frac{5-5}{16-5} = 0$$

(b)

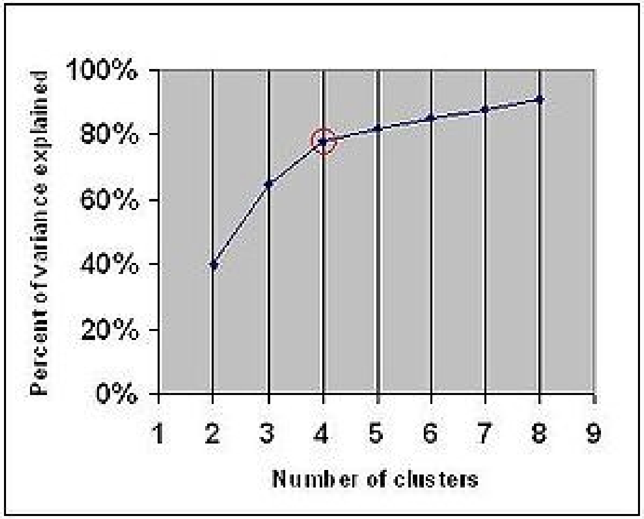
$$S = (1 + 2 + 3) + 1 = 7$$

$$C_{(b)} = \frac{7-5}{16-5}$$
$$= \frac{2}{11} > 0$$

(C-index → smaller the better)

(cont'd)

- **"External" quality metrics**
  - An alternative approach can be applied if we do in fact have labels for the data (but chose not to use it during the clustering)
  - In this case, can use any of supervised measures, such as GINI impurity, Information Gain, etc..
- **Model selection**
  - Evaluation using the quality measure mentioned here
    - e.g. by evaluating each value of *k* and finding the "kink" in the metric curve (shown on right)
- **Reliable clustering**
  - Clustering algorithms like k-means (*et al*) are heuristics and may not be globally optimal.
    - By repeating clustering operations multiple times and selecting the best options we can obtain more robust clusters
  - Also possible to use optimization algorithms like GA and Particle Swarm to directly optimize these quality metrics



Fuzzy C-Means

- ***K*-means algorithms → severe local minimum problems**
  - Standard algorithm is very “unforgiving”
    - Sensitivity to initial conditions
    - Clusters can get “marginalized” very easily
- **Previously discussed the “mixture of gaussians” technique**
  - Allows points to belong to >1 clusters, more elegant, etc.
  - In general, big improvement over *k*-means
  - However, requires “EM algorithm” → mathematically complex (and not covered in this course :-))
- **Can we find a compromise?**

  → ***Fuzzy C-Means algorithm***
  - Essentially “fuzzified” form of k-means
  - Proposed in 1981

(Cont'd)

- **General idea:**
  - Replace idea of “crisp” partitioning of clusters with set of weights

    (in fuzzy logic parlance, this is the “membership function”
  - Each point assigned a set of weights $w_{i,j}$ → which is the membership of point *i* in cluster *j*

$$\sum_{j=1}^{k} w_{i,j} = 1$$

  - i.e. all weights for a point sum to 1
  - Each cluster c_j should contain at least one point *i* for which $w_{i,j}$ is non-zero.

Modified algorithm

1. Randomly select initial weights
2. Repeat until convergence criteria met:
   i. Compute centroid of each cluster using:

$$c_j = \frac{\sum_{i=1}^{n} w_{ij}^{p} x_i}{\sum_{i=1}^{n} w_{ij}^{p}}$$

   ii. Update weights/membership functions using:

$$w_{ij} = \frac{(1/\delta_{ij}^2)^{\frac{1}{p-1}}}{\sum_{q=1}^{k} (1/\delta_{iq}^2)^{\frac{1}{p-1}}}$$

p=1 → k-means

p>1 → increasing levels of “fuzziness”