Unsupervised Learning

A review of clustering and other exploratory data analysis methods

Harvard-MIT Division of Health Sciences and Technology
HST.951J: Medical Decision Support

A few “synonyms”…

- Agminatics
- Aciniformics
- Q-analysis
- Botryology
- Systematics
- Taximetrics
- Clumping
- Morphometrics
- Nosography
- Nosology
- Numerical taxonomy
- Typology
- Clustering

- A multidimensional space needs to be reduced…

What we are trying to do

Predict this

	age	test1	
Case 1	0.7	-0.2	0.8
Case 2	0.6	0.5	-0.4
	-0.6	0.1	0.2
	0	-0.9	0.3
	-0.4	0.4	0.2
	-0.8	0.6	0.3
	0.5	-0.7	-0.4

Using these

We are trying to see whether there seems to exist patterns in the data…

Exploratory Data Analysis

- Hypothesis generation versus hypothesis testing…
- The goal is to visualize patterns and then interpret them

- Unsupervised: No GOLD STANDARD

See Khan et al. Nature Medicine, 7(6): 673 - 679.

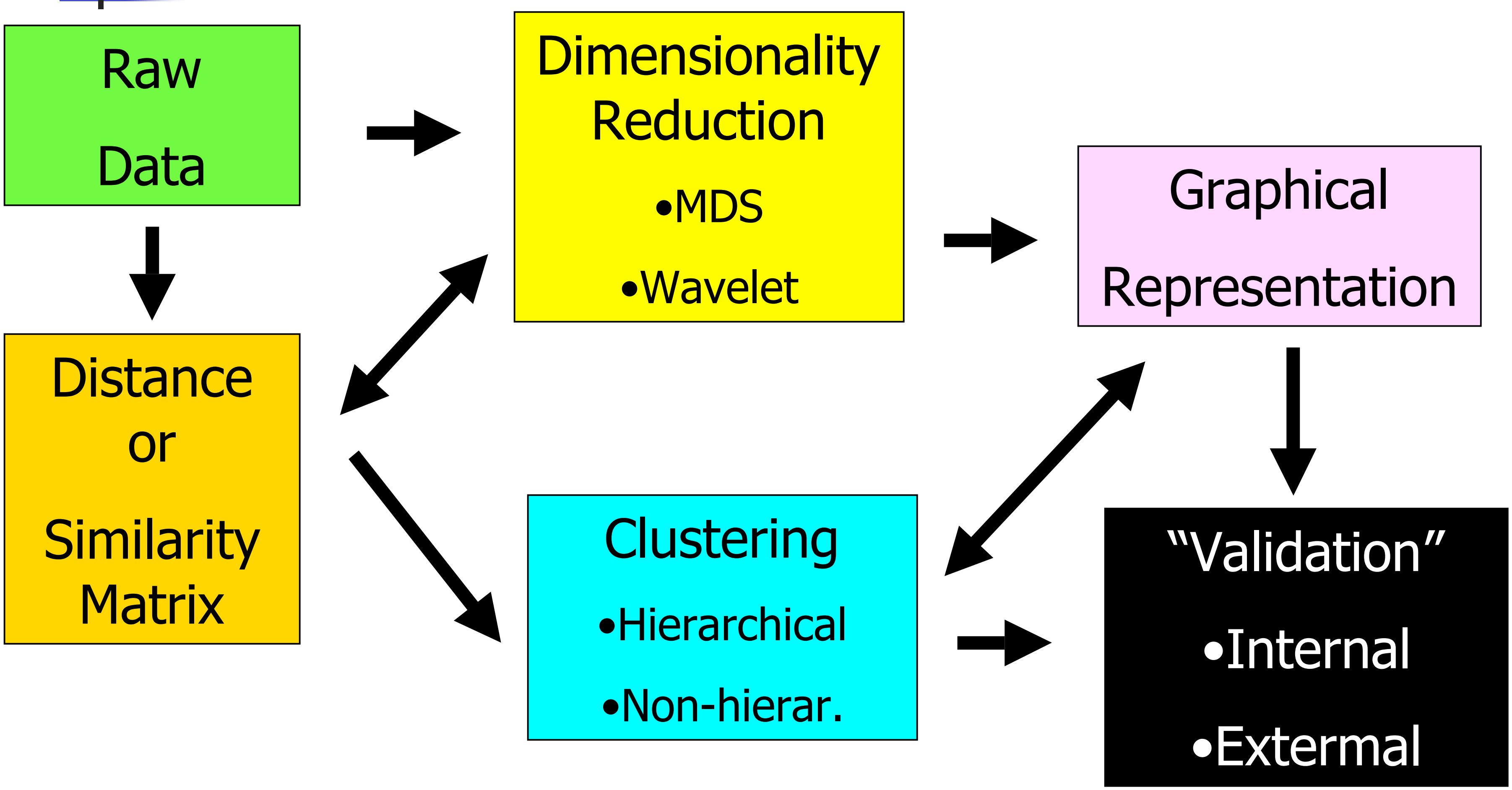
Outline

- Proximity
  - Distance Metrics
  - Similarity Measures
- Clustering
  - Hierarchical Clustering
    - Agglomerative
  - K-means
- Multidimensional Scaling
- Graphical Representations

Similarity between objects

Similarity Data	Relation of Data to Spatial Representation	Spatial Representation
Percent “same” judgments for all pairs of successively presented aural signals of the International Morse Code (see Rothkopf, 1957).	Obtained relation between Rothkopf’s original similarity data for the 36 Morse Code signals and the Euclidean distances in Shepard’s spatial solution.	Two-dimensional spatial solution for the 36 Morse Code signals obtained by Shepard (1963) on the basis of Rothkopf’s (1957) data.

Unsupervised Learning



Algorithms, similarity measures, and graphical representations

- Most algorithms are not necessarily linked to a particular metric or similarity measure
- Also not necessarily linked to a particular graphical representation

- There has been interest in this given high throughput gene expression technologies
- Old algorithms have been rediscovered and renamed

Metrics

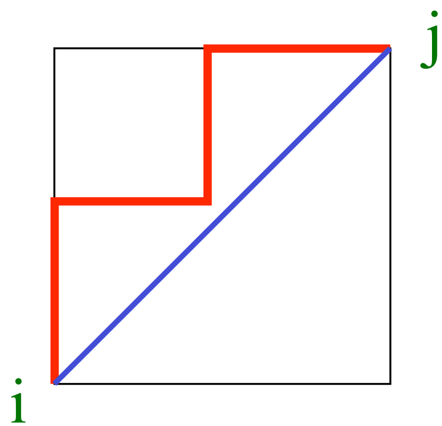
Minkowski r-metric

$$d_{ij} = \left(\sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^{r} \right)^{1/r}$$

- Manhattan
  - (city-block)

$$d_{ij} = \left(\sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right| \right)$$

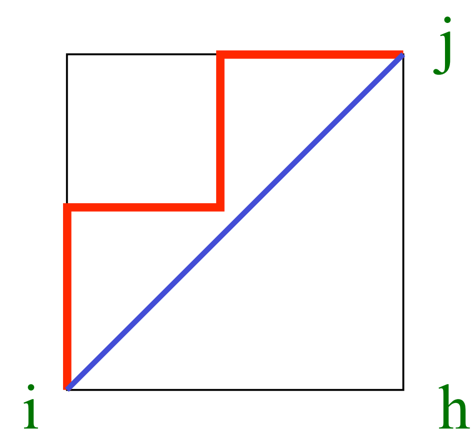
- Euclidean



$$d_{ij} = \left(\sum_{k=1}^{K} \left| x_{ik} - x_{jk} \right|^{2} \right)^{1/2}$$

Metric spaces

- Positivity
  Reflexivity $$d_{ij} > d_{ii} = 0$$
- Symmetry $$d_{ij} = d_{ji}$$
- Triangle
  inequality $$d_{ij} \le d_{ih} + d_{hj}$$

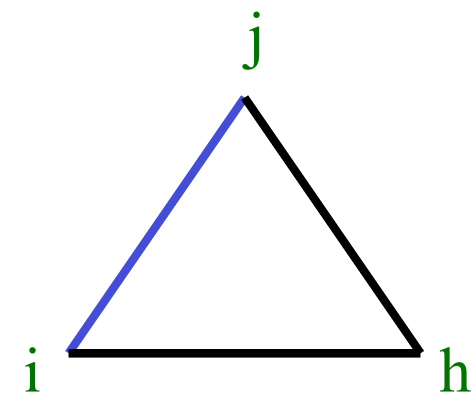


More metrics

- Ultrametric $d_{ij} \leq \max\left[d_{ih}, d_{hj} \right]$

  replaces

  $d_{ij} \leq d_{ih} + d_{hj}$



- Four-point additive condition $d_{hi} + d_{jk} \leq \max\left[(d_{hj} + d_{ik}), (d_{hk} + d_{ij}) \right]$

  replaces

  $d_{ij} \leq d_{ih} + d_{hj}$

Similarity measures

- Similarity function
  - For binary, “shared attributes”

$$s(i,j) = \frac{i^t j}{\|i\| \|j\|}$$

$$s(i,j) = \frac{1}{\sqrt{2 \square 1}}$$

$$i^t = [1,0,1]$$

$$j = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Variations…

- Fraction of d attributes shared

$$s(i,j) = \frac{i^t j}{d}$$

- Tanimoto coefficient

$$s(i,j) = \frac{i^t j}{i^t i + j^t j - i^t j}$$

$$s(i,j) = \frac{1}{2 + 1 - 1}$$

$$i^t = [1,0,1]$$

$$j = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

More variations…

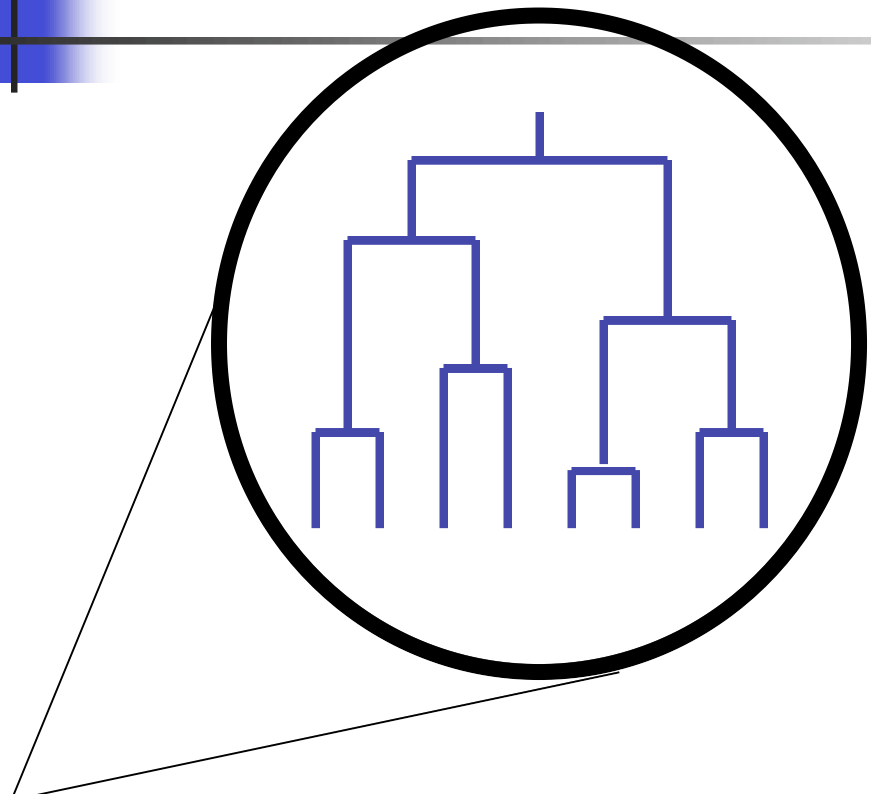
- Correlation
  - Linear
  - Rank
- Entropy-based
  - Mutual information
- Ad-hoc
  - Neural networks

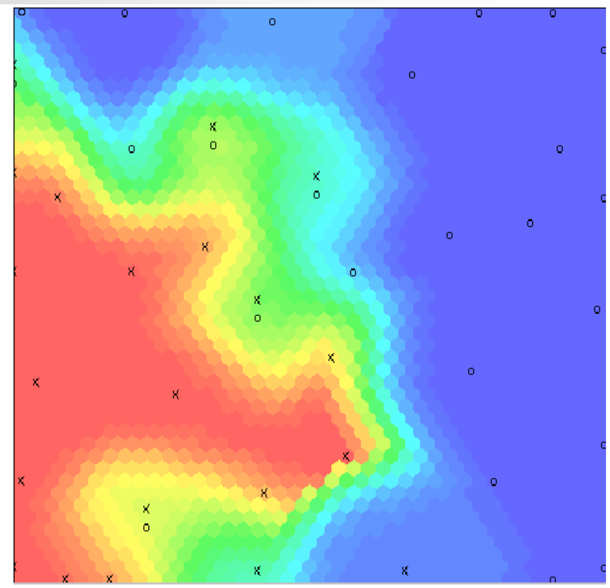
Clustering

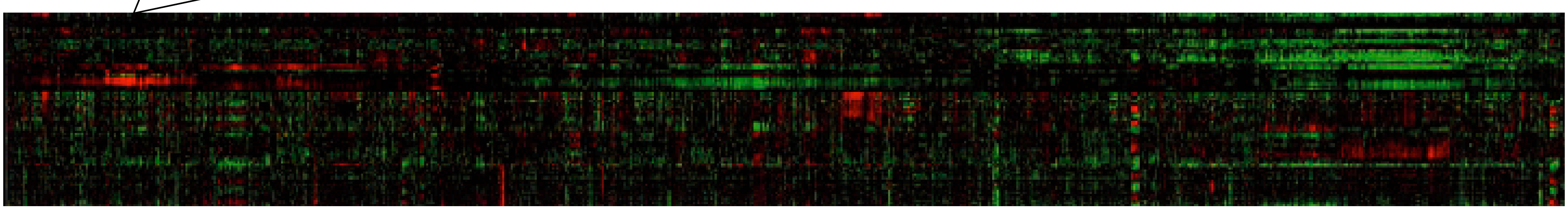
Hierarchical Clustering

- Agglomerative Technique
  - Successive “fusing” cases
  - Respect (or not) definitions of intra- and /or inter-group proximity
- Visualization
  - Dendrogram, Tree, Venn diagram

Data Visualization







Linkages

- Single-linkage: proximity to the closest element in another cluster
- Complete-linkage: proximity to the most distant element
- Mean: proximity to the mean (centroid)

Graphical Representations

Hierarchical

Additive Trees

- Commonly the minimum spanning tree
- Nearest neighbor approach to hierarchical clustering

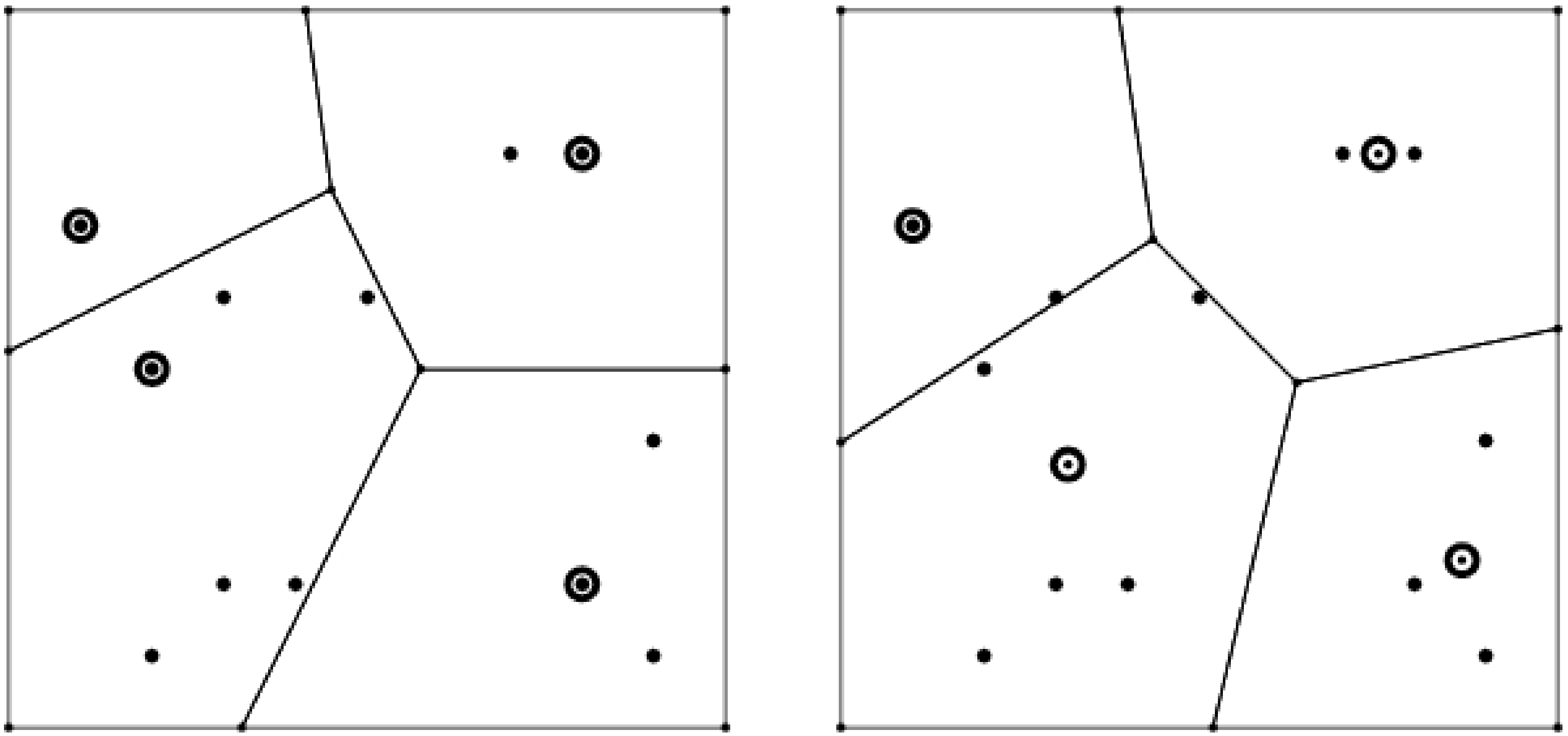
Non-Hierarchical: Distance threshold

See Duda et al., “Pattern Classification”

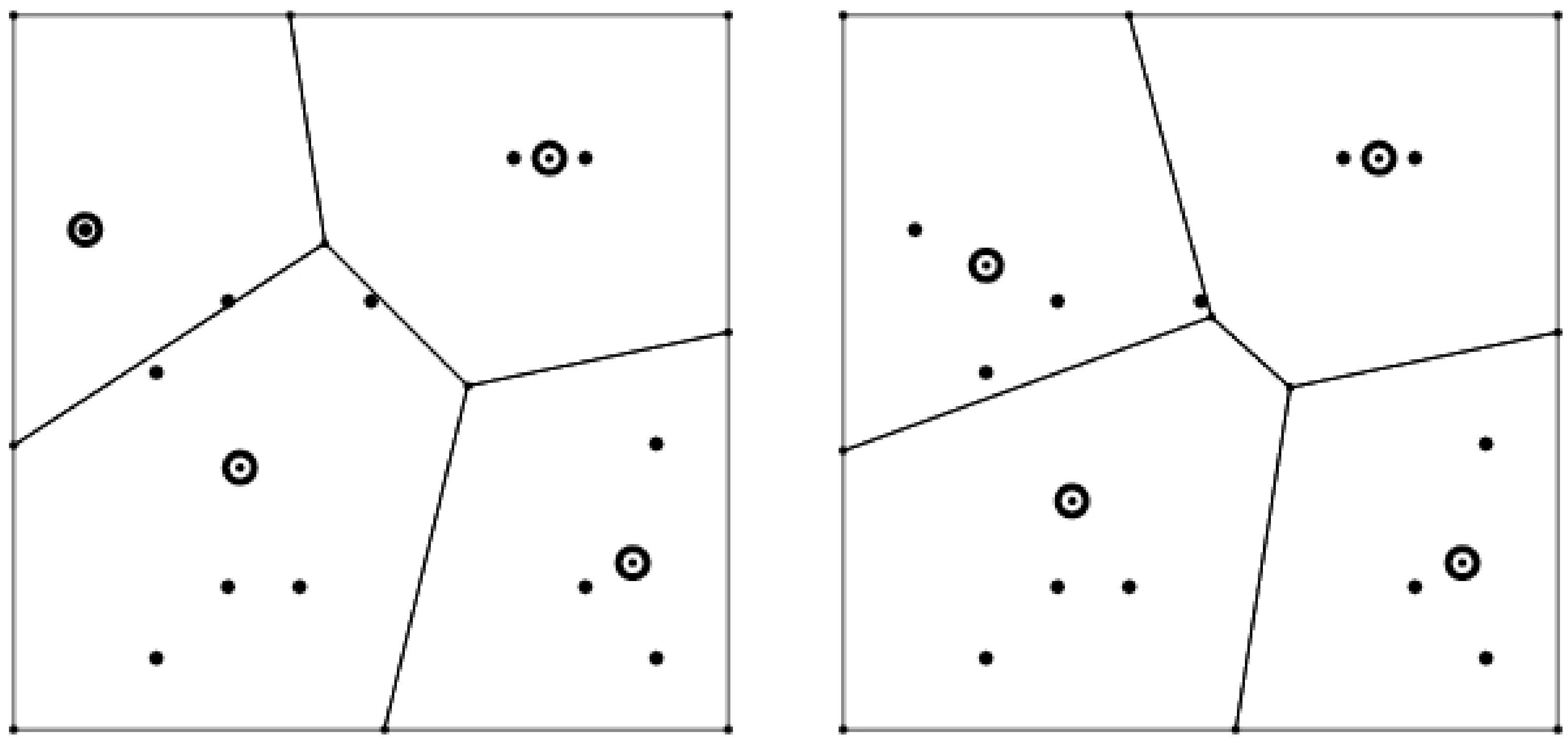
k-means clustering (Lloyd's algorithm)

1. Select k (number of clusters)
2. Select k initial cluster centers $c_1,\ldots,c_k$
3. Iterate until convergence: For each i,
    1. Determine data vectors $v_{i1},\ldots,v_{in}$ closest to c_i (i.e., partition space)
    2. Update c_i as $c_i = 1/n\,(v_{i1}+\ldots+v_{in})$

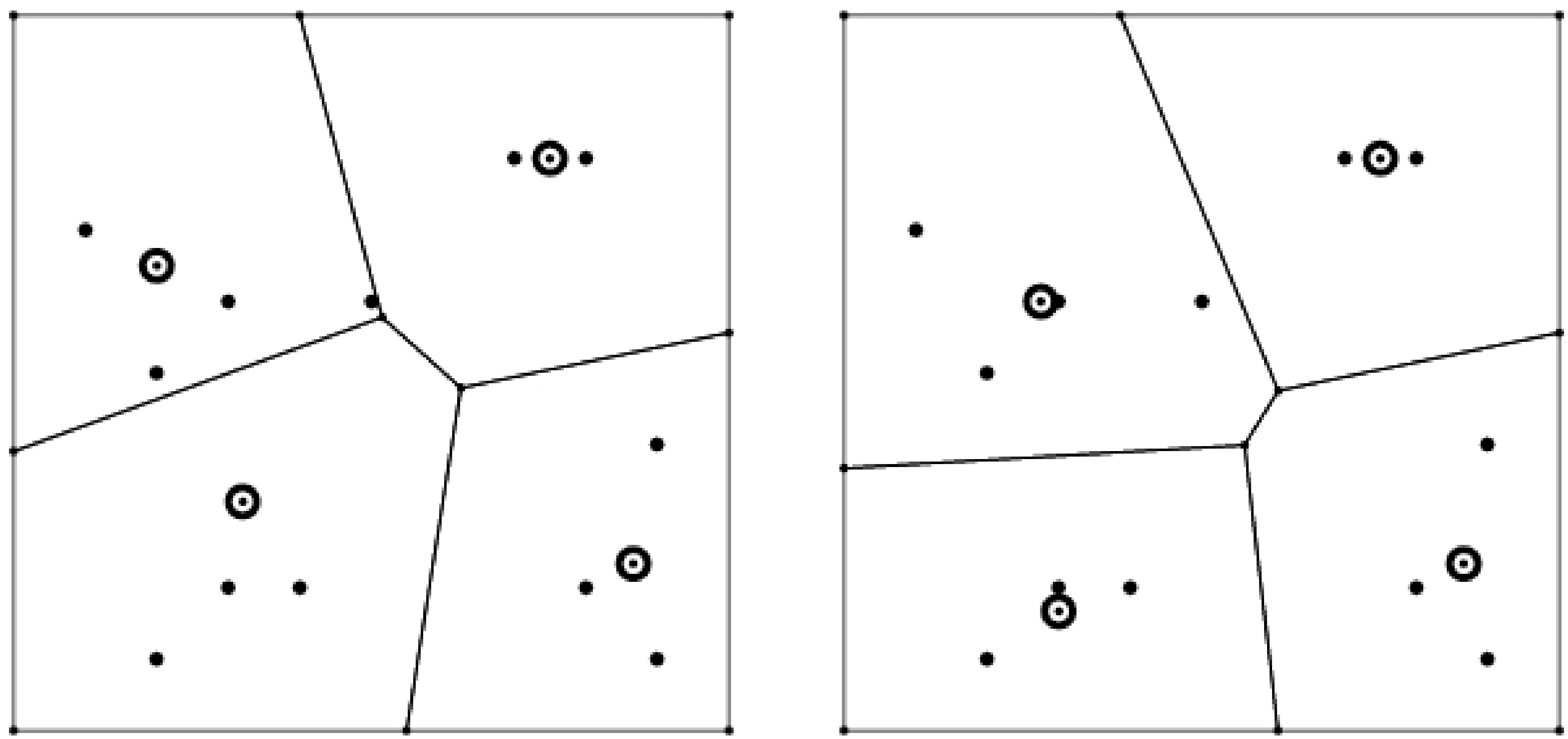
k-means clustering example



k-means clustering example



k-means clustering example



Common mistakes

- Refer to dendrograms as meaning "hierarchical clustering" in general
- Misinterpretation of tree-like graphical representations
- Ill definition of clustering criterion
  - Declare a clustering algorithm as "best"
- Expect classification model from clusters
- Expect robust results with little/poor data

Dimensionality Reduction

Multidimensional Scaling

- Geometrical models
- Uncover structure or pattern in observed proximity matrix
- Objective is to determine both dimensionality *d* and the position of points in the *d*-dimensional space

Metric and non-metric MDS

- Metric (Torgerson 1952)
- Non-metric (Shepard 1961)
  - Estimates nonlinear form of the monotonic function

$$s_{ij} = f_{mon}(d_{ij})$$

Similarity Data

Judged similaritied between 14 spectral colors varying in wavelength from 434 to 674 nanometers (from Ekman, 1954)

Relation of Data to Spatial Representation

Obtained relation between Ekman's original similarity data for the 14 colors and the Euclidean distances in Shepard's spatial solution.

Spatial Representation

Two-dimensional spatial solution for the 14 colors obtained by Shepard (1962) on the basis of Ekman's (1954) similarity data.

Stress and goodness-of-fit

Stress	Goodness of fit
20	Poor
10	Fair
5	Good
2.5	Excellent
0	Perfect

References

- Reference books for this course (Duda and Hard, Hastie et al.)
- B. Everitt
- J. Hartigan
- R. Shepard

- Sage books