Lecture 12: Clustering

COMP90049
Knowledge Technologies

Sarah Erfani and Karin Verspoor, CIS

Semester 2, 2017

THE UNIVERSITY OF MELBOURNE

Example clusters for the weather dataset

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	~~no~~
sunny	hot	high	TRUE	~~no~~
overcast	hot	high	FALSE	~~yes~~
rainy	mild	high	FALSE	~~yes~~
rainy	cool	normal	FALSE	~~yes~~
rainy	cool	normal	TRUE	~~no~~
⋮	⋮	⋮	⋮	~~⋮~~

A possible clustering of the weather dataset

Outlook	Temperature	Humidity	Windy	Cluster
sunny	**hot**	**high**	FALSE	0
sunny	**hot**	**high**	TRUE	0
overcast	**hot**	**high**	FALSE	0
rainy	mild	high	FALSE	1
rainy	**cool**	**normal**	FALSE	1
rainy	**cool**	**normal**	TRUE	1
overcast	**cool**	**normal**	TRUE	1
sunny	mild	**high**	FALSE	0
sunny	**cool**	**normal**	FALSE	1
rainy	**mild**	**normal**	FALSE	1
sunny	**mild**	**normal**	TRUE	1
overcast	**mild**	high	TRUE	1
overcast	**hot**	normal	FALSE	0
rainy	**mild**	high	TRUE	1

Clustering over the weather dataset (cf. outputs)

Outlook	Temperature	Humidity	Windy	Cluster	Play
sunny	hot	high	FALSE	0	no
sunny	hot	high	TRUE	0	no
overcast	hot	high	FALSE	0	yes
rainy	mild	high	FALSE	1	yes
rainy	cool	normal	FALSE	1	yes
rainy	cool	normal	TRUE	1	no
overcast	cool	normal	TRUE	1	yes
sunny	mild	high	FALSE	0	no
sunny	cool	normal	FALSE	1	yes
rainy	mild	normal	FALSE	1	yes
sunny	mild	normal	TRUE	1	yes
overcast	mild	high	TRUE	1	yes
overcast	hot	normal	FALSE	0	yes
rainy	mild	high	TRUE	1	no

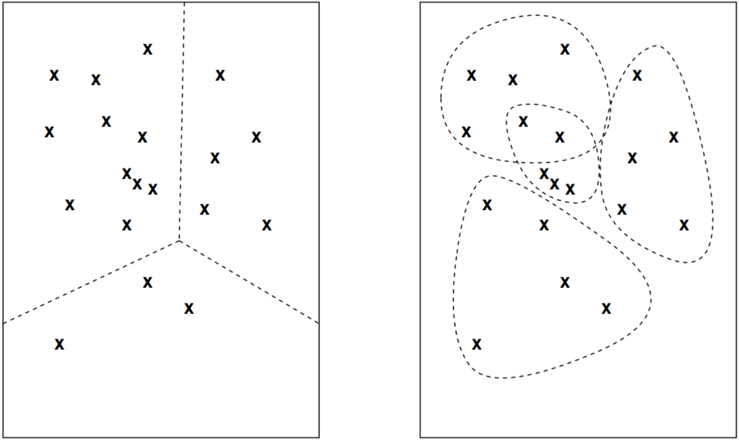
Clustering

- Clustering is ***unsupervised***
- The class of an example is not known (or at least not used)
- Finding groups of items that are *similar*
- Success often measured subjectively
- Applications in pattern recognition, spatial data analysis, medical diagnosis, . . .

Clustering, basic contrasts

- Exclusive vs. overlapping clustering
  - Can an item be in more than one cluster?
- Deterministic vs. probabilistic clustering (Hard vs. soft clustering)
  - Can an item be partially or weakly in a cluster?
- Hierarchical vs. partitioning clustering
  - Do the clusters have subset relationships between them? e.g. nested in a tree?
- Partial vs. complete
  - In some cases, we only want to cluster some of the data
- Heterogenous vs. homogenous
  - Clusters of widely different sizes, shapes, and densities
- Incremental vs. batch clustering
  - Is the whole set of items clustered in one go?

Exclusive vs. overlapping clustering

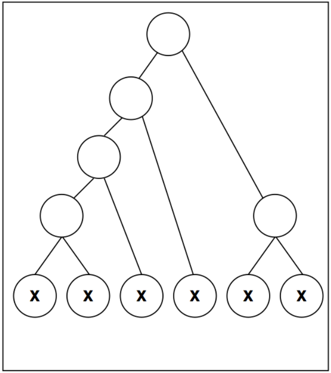


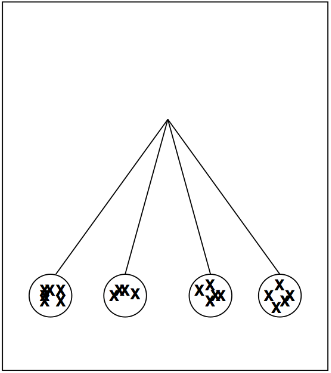
Deterministic vs. probabilistic clustering

Instance	*Cluster*
1	2
2	3
3	2
4	1
5	2
6	2
7	4
⋮	⋮

	Cluster			
Instance	1	2	3	4
1	0.01	0.87	0.12	0.00
2	0.05	0.25	0.67	0.03
3	0.00	0.98	0.02	0.00
4	0.45	0.39	0.08	0.08
5	0.01	0.99	0.00	0.00
6	0.07	0.75	0.08	0.10
7	0.23	0.10	0.20	0.47
⋮	⋮			

Hierarchical vs. partitioning clustering





Clustering, Desiderata

- Scalability; high dimensionality
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Able to deal with noise and outliers
- Insensitive to order of input records

What is a good clustering?

Two clusters?

Four clusters?

Six clusters?

Types of Evaluation

Unsupervised.

- Measures the goodness of a clustering structure without respect to external information. Includes measures of cluster cohesion (compactness, tightness), and measures of cluster separation (isolation, distinctiveness).

Supervised.

- Measures the extent to which the clustering structure discovered by a clustering algorithm matches some external structure. For instance, *entropy* can measure how well cluster labels match externally supplied class labels.

Relative.

- Compares different clusterings or clusters (using an unsupervised or supervised measure for the purpose of comparison).

Evaluating clusters mathematically

Most common measure is Sum of Squared Error (SSE) or *Scatter*

- For each point, the error is the distance to the nearest cluster
- To get SSE, we square these errors and sum them.

  $\sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x)$
- x is a data point in cluster C_i and m_i is the representative point for cluster C_i
- Can show that the m_i that minimises SSE corresponds to the center (mean) of the cluster
- Given two clusters, we can choose the one with the smallest error
- One easy way to reduce SSE is to increase k, the number of clusters
- However, a good clustering with smaller k can have a lower SSE than a poor clustering with higher k

Similarity / Proximity / Closeness

A key component of any clustering algorithm is a measurement of the distance between any points.

- Data points in Euclidean space
  - Euclidean distance
  - Manhattan (L1) distance
- Discrete values
  - Hamming distance (discrepancy between the bit strings)

d	a	b	c
a	0	1	1
b	1	0	1
c	1	1	0

    For two bit strings, the number of positions at which the corresponding symbols are different
- Documents
  - Cosine similarity
  - Jaccard measure
- Other measures
  - Correlation
  - Graph-based measures

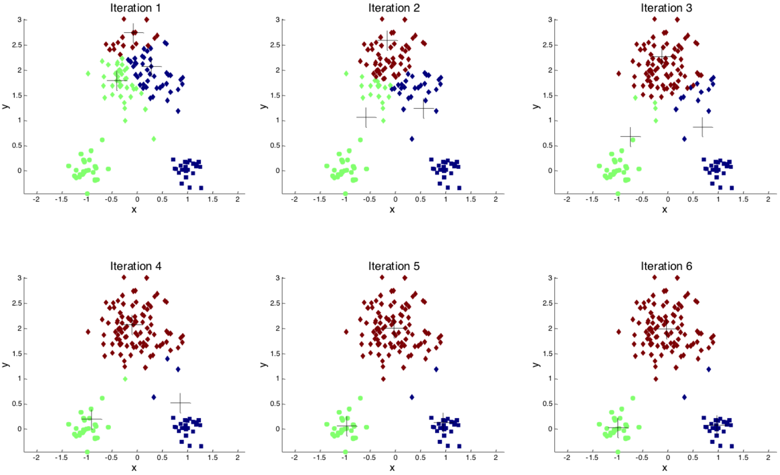
k-means Clustering

Given k, the *k*-means algorithm is implemented in four steps:

1. Select *k* points to act as seed cluster centroids
2. **repeat**
3. Assign each instance to the cluster with the **nearest centroid**
4. Recompute the centroid of each cluster
5. **until** the centroids don't change

- Exclusive, deterministic, partitioning, batch clustering method

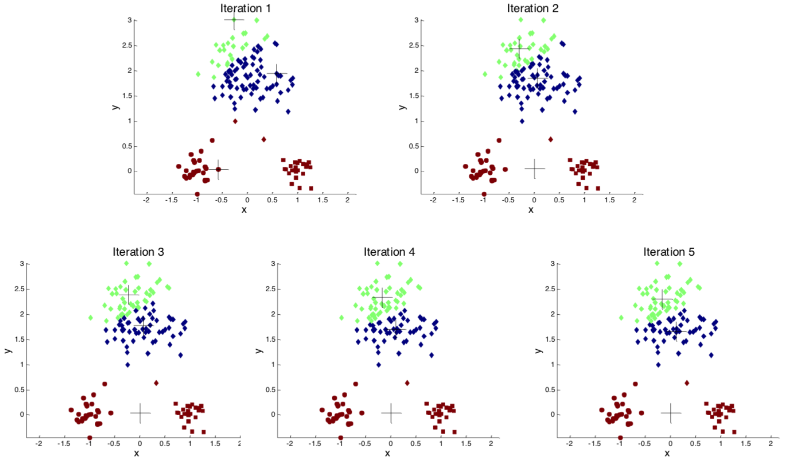
Example, Iterations



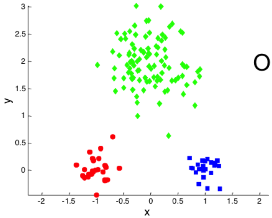
k-means Clustering – Details

- Initial centroids are often chosen randomly.
  - Clusters produced vary from one run to another.
- The centroid is (typically) the mean of the points in the cluster.
- 'Nearest' is based on proximity/similarity/etc. metric.
- K-means will converge for common similarity measures mentioned above.
  - Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters' (this way the stopping criterion will not depend on the type of similarity or dimensionality)

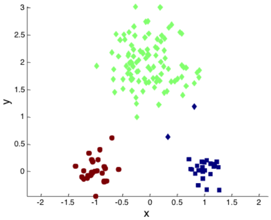
Example, Impact of initial seeds



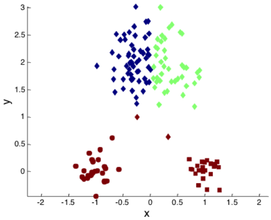
Example, Different outcomes



Original Points



Optimal Clustering



Sub-optimal Clustering

k-means, Pros and Cons

Strengths:

- relatively efficient:
  - $O(ndki)$, where n is no. instances, d is no. attributes, k is no. clusters, and i is no. iterations; normally $k, i \ll n$
  - Unfortunately we cannot a priori know the value of i!
- can be extended to hierarchical clustering

Weaknesses:

- tends to converge to local minimum; sensitive to seed instances (try multiple iterations with different seeds?)
- need to specify k in advance
- not able to handle non-convex clusters, or clusters of differing densities or sizes
- “mean” ill-defined for nominal or categorical attributes
- may not work well when the data contains outliers

Hierarchical Clustering

Bottom-up (= agglomerative) clustering

- Start with single-instance clusters
- At each step, join the two closest clusters (in terms of margin between clusters, distance between mean, ...)

Top-down (= divisive) clustering

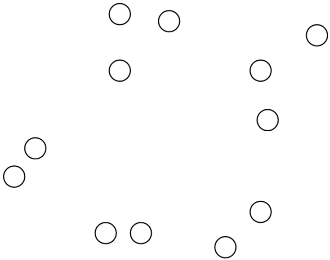
- Start with one universal cluster
- Find two partitioning clusters
- Proceed recursively on each subset
- Can be very fast

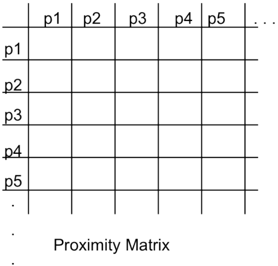
In contrast to k-means clustering, hierarchical clustering only requires a measure of similarity between *groups* of data points (no seeds, no k value).

Agglomerative Clustering

1. Compute the proximity matrix, if necessary.
2. **repeat**
3. Merge the closest two clusters
4. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters
5. **until** Only one cluster remains

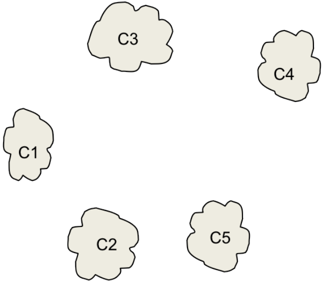
Example, Step 1





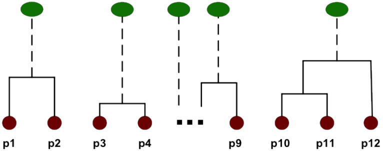
Proximity Matrix

Example, Step 2

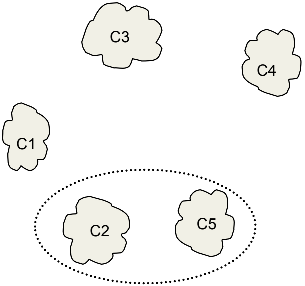


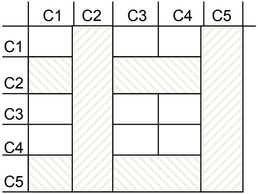
	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

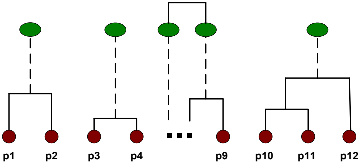


Example, Step 3





Proximity Matrix



Graph-based measure of Proximity

(a) MIN (single link.) (b) MAX (complete link.) (c) Group average.

Updating the proximity matrix:

- Single Link: *Minimum* distance between any two points in the two clusters. (most similar members)
- Complete Link: *Maximum* distance between any two points in the two clusters. (most dissimilar members)
- Group Average: *Average* distance between all points (pairwise).

Agglomerative Clustering Example

	1	2	3	4	5
1	1.00	0.90	0.10	0.65	0.20
2	0.90	1.00	0.70	0.60	0.50
3	0.10	0.70	1.00	0.40	0.30
4	0.65	0.60	0.40	1.00	0.80
5	0.20	0.50	0.30	0.80	1.00

What are the two closest points?

Agglomerative Clustering Example

	1	2	3	4	5
1	1.00	0.90	0.10	0.65	0.20
2	0.90	1.00	0.70	0.60	0.50
3	0.10	0.70	1.00	0.40	0.30
4	0.65	0.60	0.40	1.00	0.80
5	0.20	0.50	0.30	0.80	1.00

Merge points 1 & 2 into a new cluster: 6

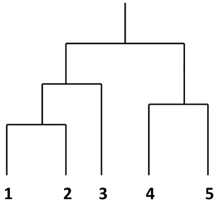
Update (single link):

	1	2	3	4	5	6
6	—	—	0.70	0.65	0.50	1.00

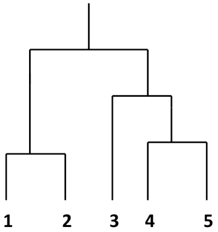
Update (complete link):

	1	2	3	4	5	6
6	—	—	0.10	0.60	0.20	1.00

	1	2	3	4	5
1	1.00	0.90	0.10	0.65	0.20
2	0.90	1.00	0.70	0.60	0.50
3	0.10	0.70	1.00	0.40	0.30
4	0.65	0.60	0.40	1.00	0.80
5	0.20	0.50	0.30	0.80	1.00



Single link



Complete link

Thoughts on Clustering

Clustering is in the eyes of the beholder

- "The validation of clustering structures is the most difficult and frustrating part of cluster analysis. Without a strong effort in this direction, cluster analysis will remain a black art accessible only to those true believers who have experience and great courage.

  – Algorithms for Clustering Data (1988) Jain and Dubes
  `http://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf`

Summary

- What basic contrasts are there in different clustering methods?
- How does k-means operate, and what are its strengths and weaknesses?
- What is hierarchical clustering, and how does it differ from partitioning clustering?
- What are some challenges we face when clustering data?

Resources:
Tan, Steinbach, Kumar (2006) Introduction to Data Mining. Chapter 8, Cluster Analysis
`http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf`

Jain, Dubes (1988) Algorithms for Clustering Data. `http://homepages.inf.ed.ac.uk/rbf/BOOKS/JAIN/Clustering_Jain_Dubes.pdf`