Lecture 13: Unsupervised Learning

COMP90049

Semester 2, 2022

Joseph West, CIS

Acknowledgement: Jeremy Nicholson, Tim Baldwin & Karin Verspoor

Roadmap

So far:

- Supervised machine learning algorithms
- Train and evaluation the performance of the classifiers

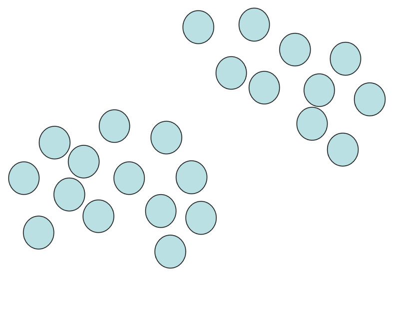
Today:

- Introduction of clustering (unsupervised learning)
- K-means clustering
- Hierarchical clustering
- Evaluate clustering performance

Introduction

What is clustering

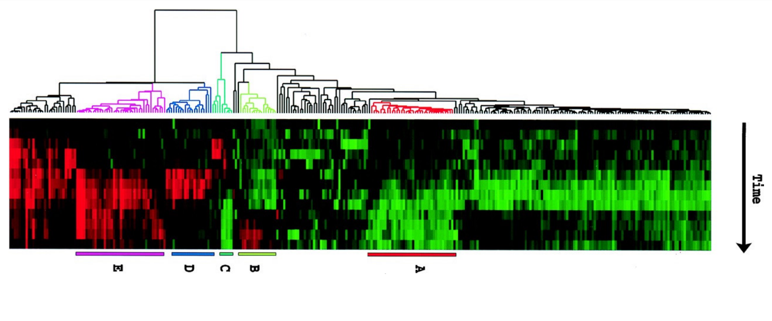
- Unsupervised learning: The class of an example is not known (or at least not used)
- Goal: find groups of similar data points
- "Similar" measure: e.g., small Euclidean distance.



Why clustering I

Applications:

- Gene expression data analysis



Eisen, M.B. et al.Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences, 95(25), 1998, pp.14863-14868.

Why clustering II

Applications:

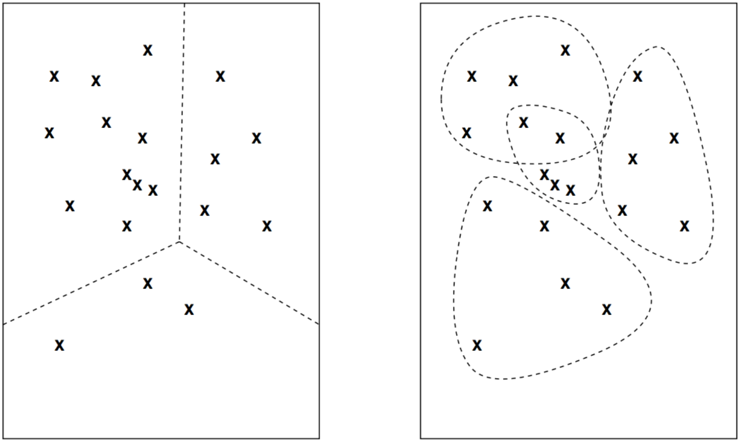
- Image segmentation

(a) original image (b) segmentation output (2 clusters) (c) segmentation output (3 clusters)

- Identify groups of customer

Exclusive vs. overlapping clustering

- Can an item be in more than one cluster?



Deterministic vs. probabilistic clustering

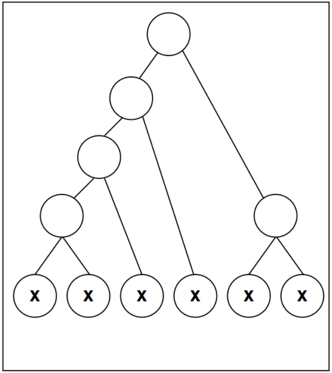
- Can an item be partially or weakly in a cluster?

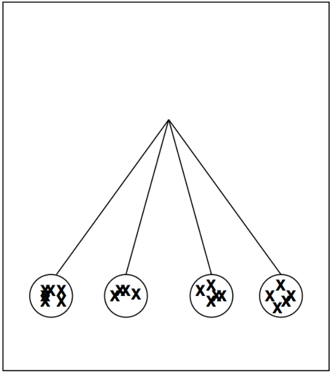
Instance	*Cluster*
1	2
2	3
3	2
4	1
5	2
6	2
7	4
⋮	⋮

	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

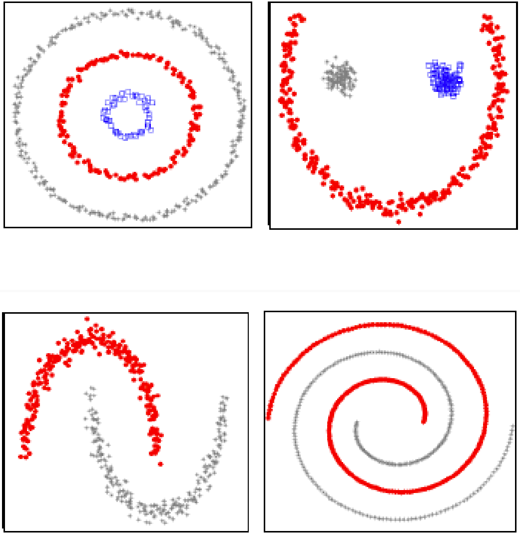
- Do the clusters have subset relationships between them? e.g. nested in a tree?





Heterogenous vs. homogenous clustering

- Clusters of widely different sizes, shapes, and densities



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?
- Heterogenous vs. homogenous
  - Clusters of widely different sizes, shapes, and densities
- Partial vs. complete clustering
  - In some cases, we only want to cluster some of the data
- Incremental vs. batch clustering
  - Is the whole set of items clustered in one go?

k-means

k-means Clustering

Given k (the number of clusters), the k-means algorithm is implemented in four steps:

1. Initialize: Select random k points to act as seed cluster centroids
2. **Iterate:**
    - Step 1 - Cluster assignment: Assign each instance to the cluster with the **nearest centroid**
    - Step 2 - Recompute the cluster centroids: Update centroid of each cluster to the average of its assigned instances
3. **Until** the centroids don't change

- Exclusive, deterministic, partitioning, batch clustering method

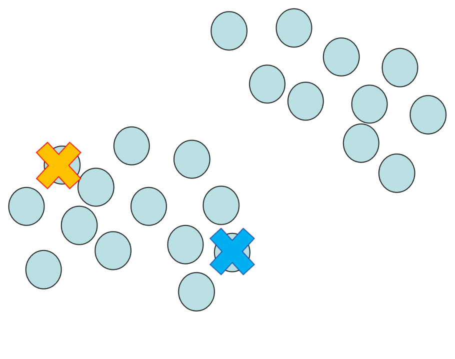
THE UNIVERSITY OF MELBOURNE

Measuring Similarity / Proximity / Closeness

- Data points in Euclidean space
  - Euclidean distance
  - Manhattan (L1) distance
- Discrete values
  - Hamming distance: discrepancy between the bit strings
- Other measures
  - Cosine similarity
  - Jaccard measure

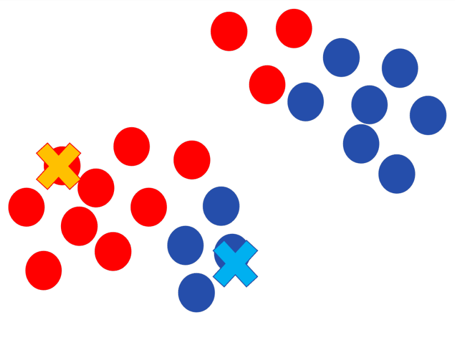
Example I

Pick 2 ($k = 2$) random points as cluster centroids.



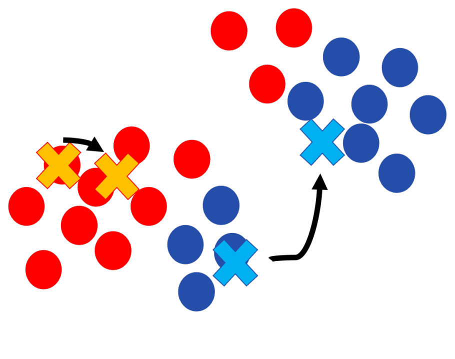
Example I

(Iteration 1) Step 1: Cluster assignment



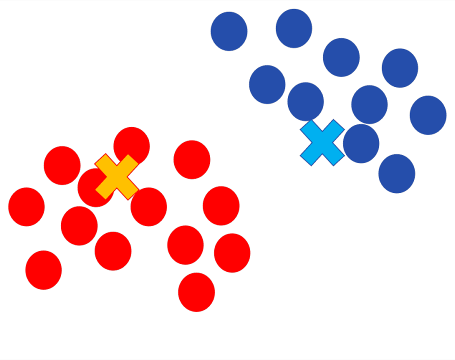
Example I

(Iteration 1) Step 2: Recompute the cluster centroids



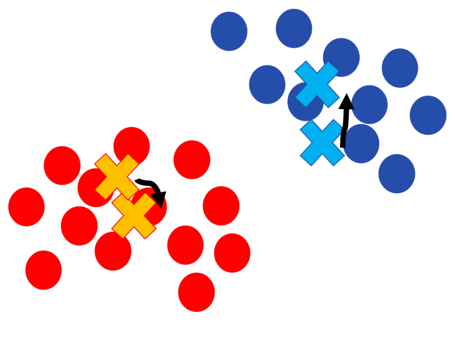
Example I

(Iteration 2) Step 1: Cluster assignment

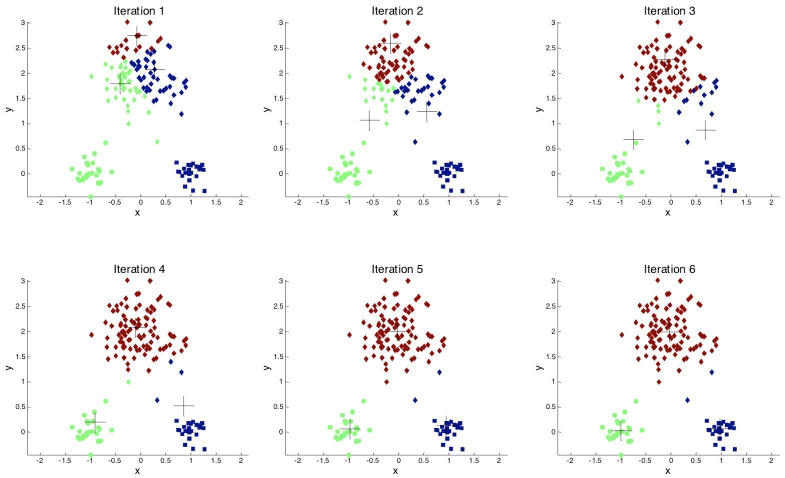


Example I

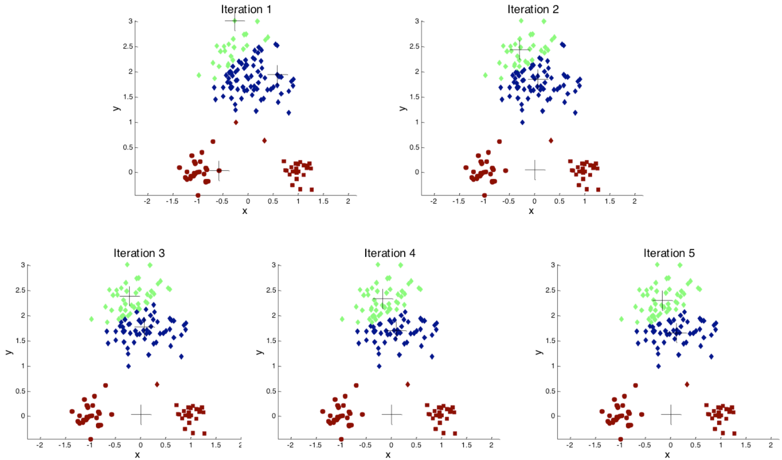
(Iteration 2) Step 2: Recompute the cluster centroids



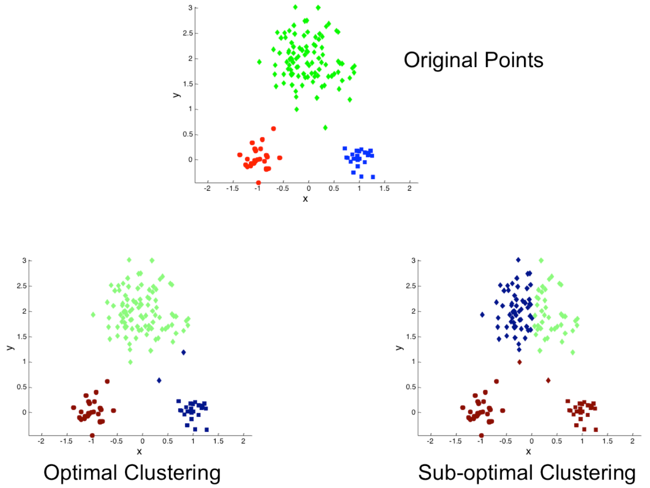
Example, Iterations



Example, Impact of initial seeds



Example, Different outcomes

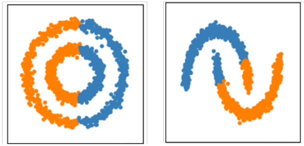


Pros of k-means

- 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

Cons of k-means

- results sensitive to random centroid selection:
  - try multiple iterations with different seeds
  - try better initialization method (k-means++)
- "mean" ill-defined for nominal or categorical attributes
- may not work well when the data contains outliers
- not able to handle non-convex clusters, or clusters of differing densities or sizes



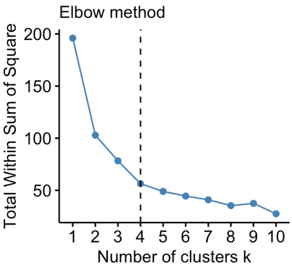
- need to specify k in advance.

How to choose the number of clusters?

- calculate within-cluster Sum of Squared Error SSE_w for different number of clusters K

$$SSE_w = \sum_{i=1}^{K} \sum_{x \in C_i} (x - m_i)^2$$

- x is a data point in cluster C_i and m_i is the centroid of cluster C_i.
- As K increases, we will have a smaller number of instances in each cluster $\rightarrow SSE_w$ decreases
- Elbow method: K increases to $K + 1$, the drop of SSE_w starts to diminish



Hierarchical Clustering

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, ...)

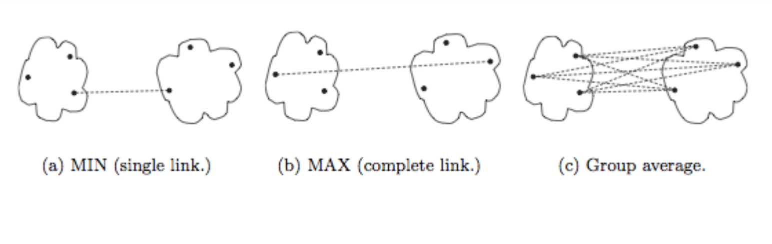
Top-down (= divisive) clustering

- Start with one universal cluster
- Find two partitioning clusters
- Proceed recursively on each subset

Bottom-up (Agglomerative) Clustering

1. Each point starts as a cluster. Compute the proximity matrix.
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

Graph-based measure of Proximity



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

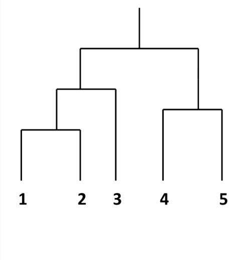
Update (single link):

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

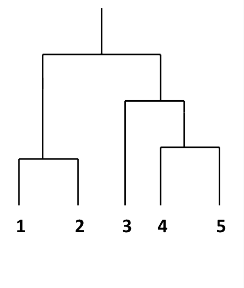
Update (complete link):

	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



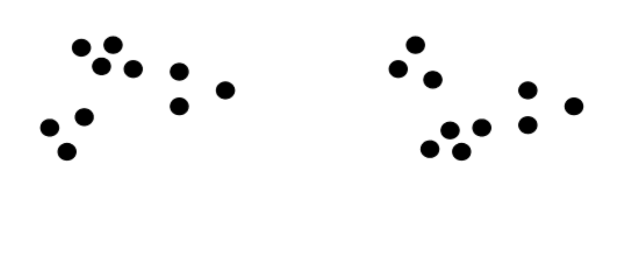
Single link



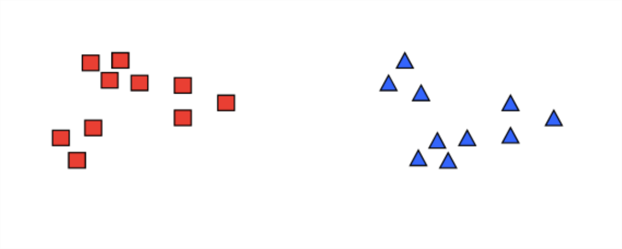
Complete link

Evaluation

What is a good clustering?

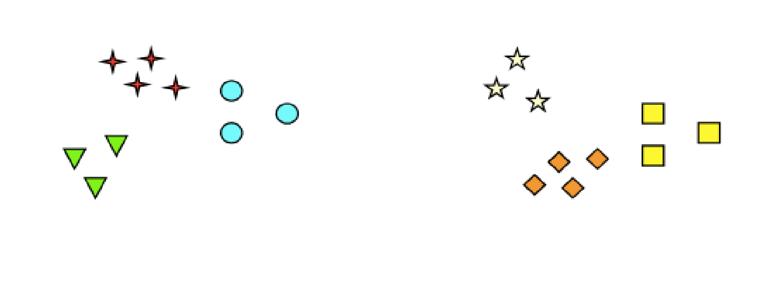


Two clusters?



Four clusters?

Six clusters?



Types of Evaluation

Unsupervised:

- cluster cohesion: compactness, tightness
- cluster separation: isolation, distinctiveness.

Supervised: measure how well cluster labels match externally supplied class labels.

- entropy
- purity

THE UNIVERSITY OF MELBOURNE

Unsupervised Evaluation I

A "good" cluster should have one or both of:

- **High Cluster Cohesion**: instances in a given cluster should be closely related to each other
- **High Cluster Separation** instances in different clusters should be distinct from each other

Unsupervised Evaluation II

Within-cluster Sum of Squared Error SSE_w: the smaller, the better

$$SSE_w = \sum_{i=1}^{K} \sum_{x \in C_i} (x - m_i)^2$$

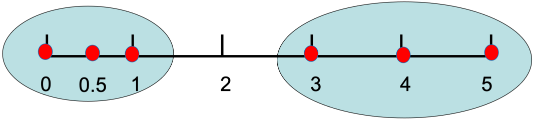
- x: a data point in cluster C_i
- m_i: the centroid for cluster C_i.

Between-cluster Sum of Squared Error SSE_b: the larger, the better

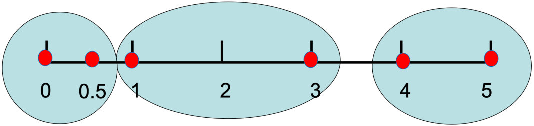
$$SSE_b = \sum_{i=1}^{K} n_i (m_i - m)^2$$

- m: mean of all data points in the dataset
- m_i: centroid for cluster C_i.
- n_i: number of instances in cluster C_i.

Example



$$SSE_w = 2.5, SSE_b = 18.375$$



$$SSE_w = ?, SSE_b = ?$$

Entropy: the smaller, the better

$$entropy = \sum_{i=1}^{k} \frac{|n_i|}{N} H_i$$

Purity: the larger, the better

$$purity = \sum_{i=1}^{k} \frac{|n_i|}{N} \max_j P_i(j)$$

- n_i: number of instances in cluster C_i.
- N: total number of instances in all clusters.
- $P_i(j)$: probability of the class j in the cluster i.
- H_i: entropy of class distribution in cluster i.

$$H_i = -\sum_j P_i(j) log_2 P_i(j)$$

Supervised Evaluation Example I

- Calculate the entropy and purity of the following cluster output

Cluster	Play = yes	Play = no
1	4	0
2	4	4

$$entropy_1 = -1 \times log(1) - 0 \times log(0) = 0$$

$$entropy_2 = -0.5 \times log(0.5) - 0.5 \times log(0.5) = 1$$

$$purity_1 = max(1, 0) = 1$$

$$purity_2 = max(0.5, 0.5) = 0.5$$

$$entropy = \frac{4}{12} \times 0 + \frac{8}{12} \times 1 = 0.67$$

$$purity = \frac{4}{12} \times 1 + \frac{8}{12} \times 0.5 = 0.67$$

Supervised Evaluation Example II

- Calculate the entropy and purity of the following cluster output

Cluster	Play = yes	Play = no
1	2	0
2	6	4

$$entropy = ?$$

$$purity = ?$$

$$entropy_1 = -1 \times log(1) - 0 \times log(0) = 0$$
$$entropy_2 = -0.6 \times log(0.6) - 0.4 \times log(0.4) = 0.97$$
$$purity_1 = max(1, 0) = 1$$
$$purity_2 = max(0.6, 0.4) = 0.6$$
$$entropy = \frac{2}{12} \times 0 + \frac{10}{12} \times 0.97 = 0.80$$
$$purity = \frac{2}{12} \times 1 + \frac{10}{12} \times 0.6 = 0.67$$

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?
- How to evaluation the clustering performance?

THE UNIVERSITY OF MELBOURNE

References

- 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*