Data Mining and Knowledge Discovery:

# Unsupervised Learning – Clustering

Bernard Ženko
(thanks to Petra Kralj Novak for the slides)

Jožef Stefan International Postgraduate School, 2021/22



## Outline

- Unsupervised learning and clustering
- Motivation and application areas
- Different approaches:
  - *k*-means clustering
  - agglomerative clustering
  - DBSCAN clustering

# Supervised vs. unsupervised

| Person | Age | Spect. presc. | Astigmatism | Tear prod. | Lenses |
|---|---|---|---|---|---|
| O1 | 17 | myope | no | reduced | none |
| O2 | 23 | myope | no | normal | soft |
| O3 | 22 | myope | yes | reduced | none |
| O4 | 27 | myope | yes | normal | hard |
| O5 | 19 | hypermetrope | no | reduced | none |
| O6 … O13 | … | … | … | … | … |
| O14 | 35 | hypermetrope | no | normal | soft |
| O15 | 43 | hypermetrope | yes | reduced | none |
| O16 | 39 | hypermetrope | yes | normal | none |
| O17 | 54 | myope | no | reduced | none |
| O18 | 62 | myope | no | normal | none |
| O19 … O23 | … | … | … | … | … |
| O24 | 56 | hypermetrope | yes | normal | none |

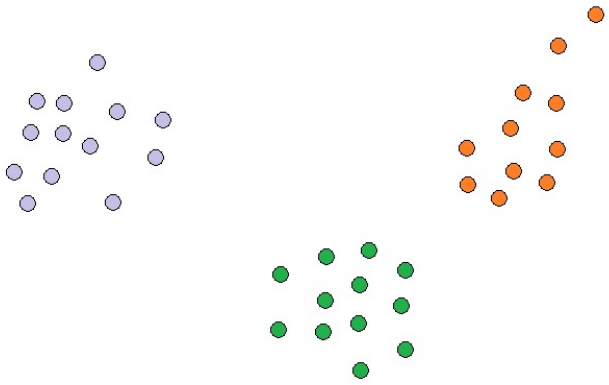
- Attribute-value representation
- Attributes, also descriptive or independent variables
- Class, also target or dependent variable
- Unsupervised learning = we have no target variable



# Unsupervised learning

- Methods:
  - clustering (cluster analysis)
  - association rules (frequent itemset mining)
  - principal component analysis (PCA)
  - …
- Exploratory data analysis, search for patterns in the data without knowing in advance what we are looking for
- Validation of results can be problematic, we typically have no objective validation measure
- Can be used (as one of the tools) in the analysis process leading to predictive models

# Clustering





# Clustering

- … is grouping of data instances into clusters so that:
  - objects within a cluster are similar to each other, and
  - dissimilar to objects in other clusters
- (Low intra-cluster variance and high inter-cluster variance)

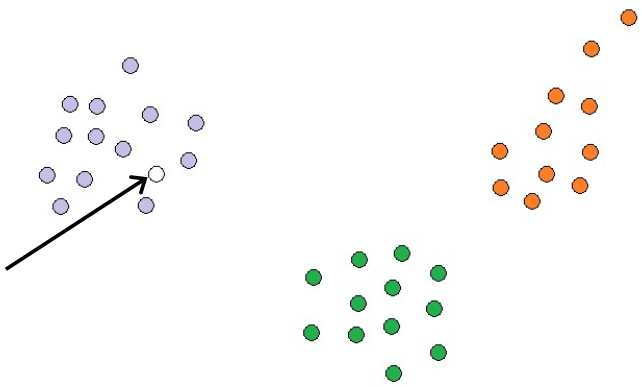
- Wish list:
  - Identification of clusters irrespective of their shapes
  - Scalability to large amounts of data
  - Ability to deal with noisy data
  - ...

# Why use clustering?

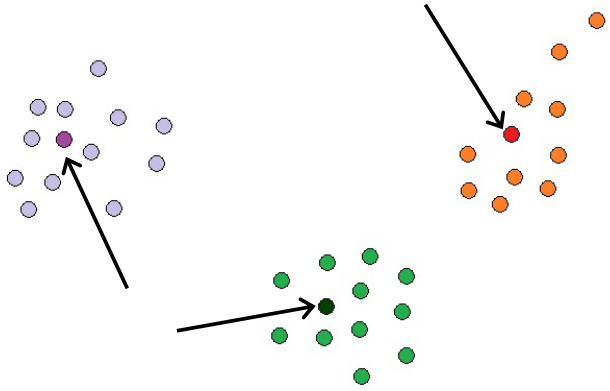
- Exploratory analysis: to detect and investigate possible patterns and similar data instances
- Unsupervised classification (conceptual clustering)
- Data summarization
- Outlier detection and analysis



# Unsupervised classification



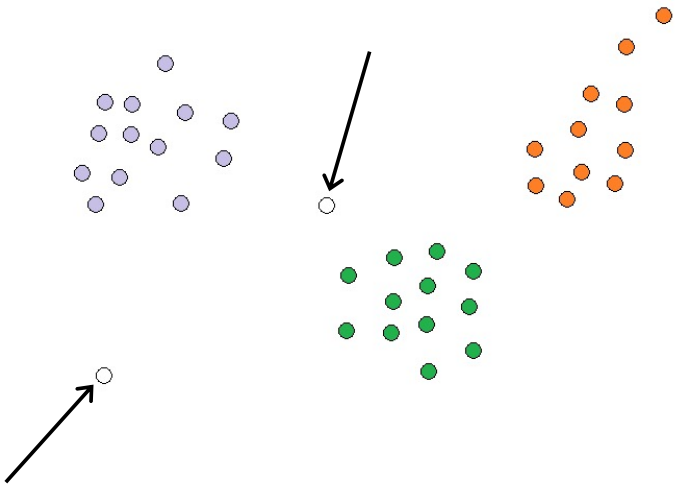
# Data summarization



- Centroid: geometric center, arithmetic mean position of all data points
- Medoid: data point for which the sum of dissimilarities to all other points in the cluster is minimal; its value is restricted to the data set (domain) values



# Outlier detection



# Applications

- Customer segmentation
- Market basket analysis
- Recommender systems
- Text mining applications
- Social network analysis
- ...

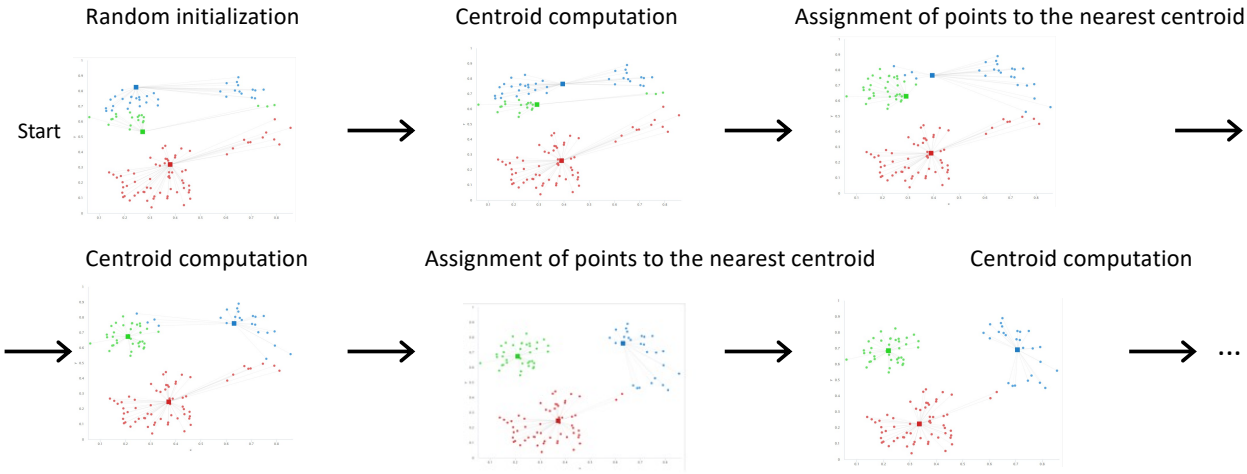


# Clustering approaches

- Partitioning clustering:
  - *k*-means, k-medoids, k-modes
- Hierarchical clustering:
  - Agglomerative clustering
- Grid-based clustering:
  - Multi-resolution grid structure
  - Efficient and scalable
- Density-based clustering:
  - A cluster is a dense region of points, which is separated by low density regions, from other regions of high density
  - Algorithms: DBSCAN, OPTICS, DenClue

# *k*-means – example

Interactive k-Means (Education






# *k*-means algorithm

1. Choose *k* random instances as cluster centers
2. Assign each instance to its closest cluster center
3. Re-compute cluster centers by computing the mean (*centroid*) of the instances assigned to each cluster
4. If cluster centers have moved, go back to step 2

- Equivalent stopping criterion: when assignment of instances to cluster centers has not changed
- Alternative cluster summarization: *k*-medoids, *k*-modes

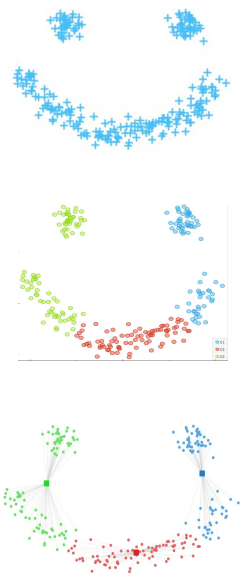
# *k*-means' shortcomings

- *k* needs to be known, but typically is not:
  - we can try different values and select the one that gives the best result according to some measure, such as *silhouette coefficient*
- Sensitivity to initial cluster center selection:
  - algorithm might get stuck in local minima
  - we can rerun the algorithm many times
  - *k*-means++: 'smart' instead of random initialization



# *k*-means' properties

- The number of clusters *k* is fixed in advance
- Is fast, always converges
- Can converge into a local minima (bad solution because of unlucky start)
- Finds convex ('spherical') shaped clusters
- *k*-means will cluster the data even if it can’t be clustered (e.g. data that comes from uniform distributions)



# Clustering evaluation

- There is no solid, objective evaluation metric

- External validation criteria:
  - use the ground truth for evaluation (if it exists)
  - use background or expert knowledge
- Internal validation criteria:
  - sum of distances to centroids
  - intra-cluster to inter-cluster distance ratio
  - silhouette coefficient
  - elbow method: plot the explained variance ($R^2$) for different *k*s and select one at the elbow
  - ...

- More: Aggarwal, Charu C. Data mining: the textbook. Springer, 2015. Ch. 6: cluster analysis, pp. 195–201, or some other textbook, or Wikipedia, or ...



# Silhouette coefficient

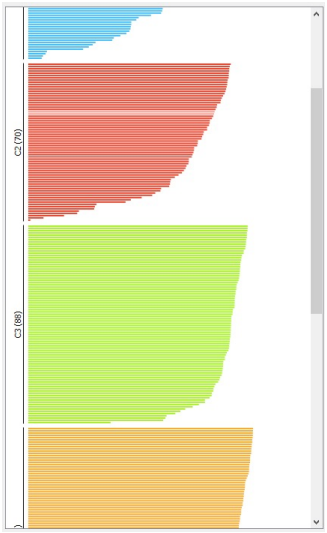
- The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation)

- For object $x_i$ , its silhouette coefficient is $s_i = (b_i - a_i)/\max(a_i, b_i)$
  - $a_i$ average distance between $x_i$ and all other objects in its cluster
  - $b_i$ average distance between $x_i$ and objects in the "neighboring" cluster, i.e., the cluster for which $b_i$ has the smallest value

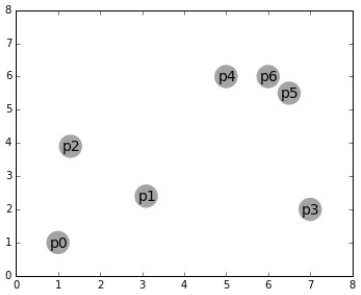
- $s_i \in [-1,1]$, we wish for $a_i > b_i$ and $a_i$ close to 0

- Silhouette coefficient of a cluster is the average of coefficients of the cluster elements
- Silhouette coefficient of a clustering is the average of coefficients of all elements

- More: Tan P-N, Kumar, 2006, Introduction to Data Mining, Ch. 8, https://www-users.cs.umn.edu/~kumar001/dmbook/ch8.pdf

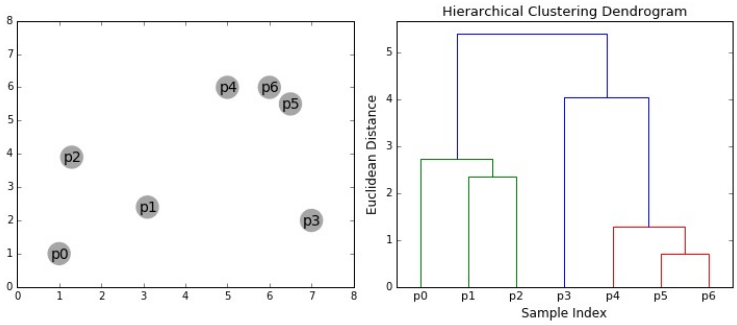


# Agglomerative clustering – example






# Agglomerative clustering – dendrogram



# Agglomerative clustering – algorithm

1. Start with a collection $C$ of $n$ singleton clusters (each cluster contains one object $c_i = \{x_i\}$ )
2. Repeat until only one cluster is left:
   1. Find a pair of clusters that is closest: min $D(c_i, c_j)$
   2. Merge the clusters $c_i$ and $c_j$ into $c_{i+j}$
   3. Remove $c_i$ and $c_j$ from the collection $C$, add $c_{i+j}$

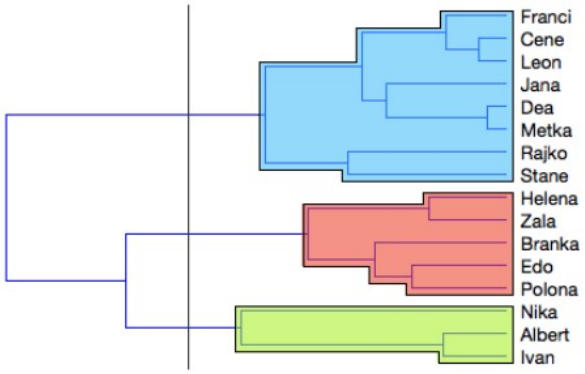
new cluster index, not a sum

- Parameters:
  - distance measure
  - linkage type (how distance between two clusters is computed)
- Shortcomings:
  - time and space complexity
  - sensitive to noisy data



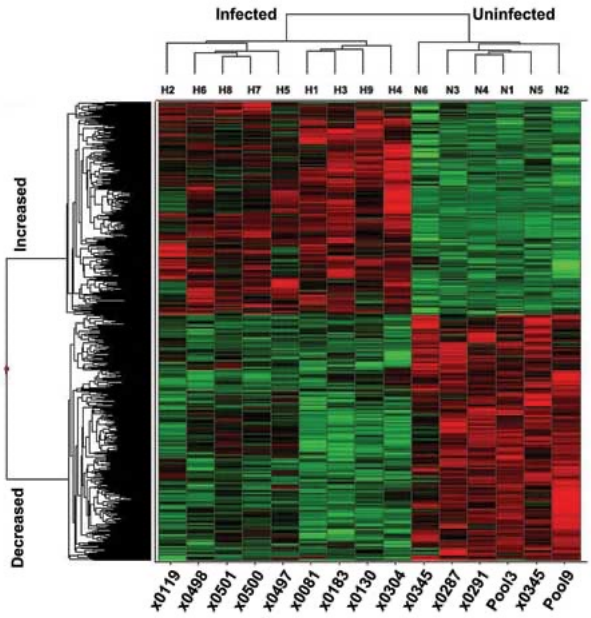
# Dendrogram

- Algorithm produces a full hierarchy of clusters visualized as a dendrogram
- Tree diagram showing (taxonomic) cluster relationships
- Can be used for selecting the number of clusters ($x$-axis can present some measure of cluster similarity)



# Example application

- Hierarchical clustering of genes

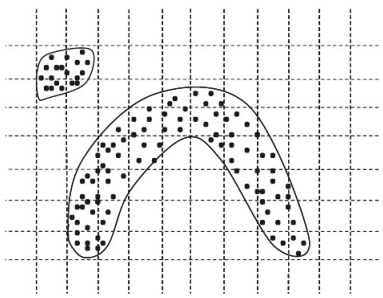


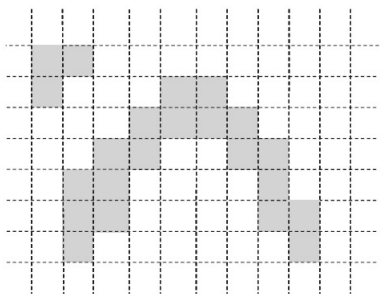



# Grid-based clustering – basic principle

1. Discretize each dimension (attribute) into $p$ ranges
2. Determine dense grid cells (at some level $\tau$, that can change)
3. Create a graph where dense grid cells are connected if they are adjacent
4. Determine connected components of graph
5. Return: Clusters represented as points in each connected component

- $p$ and $\tau$ are parameters
- fast, efficient
- can be parallelized

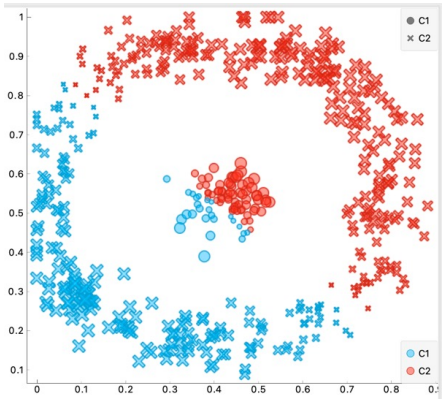




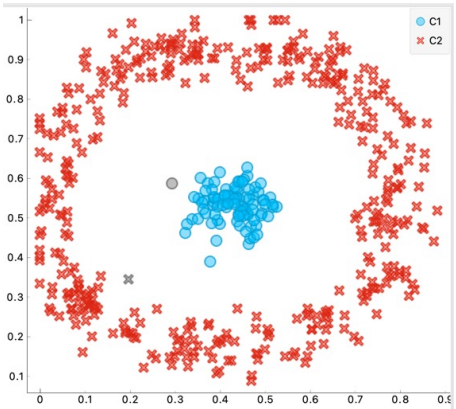
# Density-based clustering – basic principle

- Identify high-density and low-density data space regions
- Find clusters, which are contiguous high-density regions, separated from other clusters by contiguous low-density regions

*k*-means, *k*=2




DBSCAN



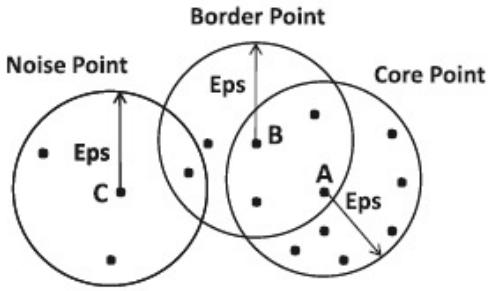



# DBSCAN clustering

= Density-Based Spatial Clustering of Applications with Noise

Definitions:

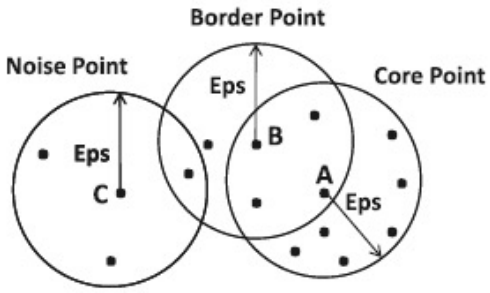
- Core point:
  - contains at least $\tau$ data points within a radius *Eps*
- Border point:
  - not a core point
  - at least one core point within a radius *Eps*
- Noise point:
  - neither a core point nor a border point




- density $\tau$ and radius *Eps* are parameters

# DBSCAN – algorithm

1. Determine core, border and noise points given (*Eps, τ*)
2. Create a graph in which core points are connected, if they are within *Eps* of one another
3. Determine connected components in the graph
4. Assign each border point to connected component with which it is best connected
5. Return points in each connected component as a cluster




- More: Aggarwal, Charu C. Data mining: the textbook. Springer, 2015. Ch. 6: cluster analysis, pp. 183, or some other textbook, or Wikipedia, or ...



# DBSCAN – properties

- Similar to grid-based approaches, except that it uses circular regions as building blocks
- Advantages:
  - can detect clusters of arbitrary shape
  - does not require the number of clusters as an input parameter
  - not sensitive to outliers
- Disadvantages:
  - computationally expensive in the first step (determining core, border and noise points)
  - susceptible to variations in the local cluster density
  - struggles with high-dimensionality data

# To be continued …