HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Spatial Data Mining
Clustering

Antti Leino ‹antti.leino@cs.helsinki.fi›

Department of Computer Science

Clustering

- Divide the data into clusters, so that
  - Points in the same cluster as similar as possible
  - Points in different clusters as different as possible
- Ancient and venerable topic in statistics
  - Plenty of clustering algorithms
  - Typically used to cluster observations
- In spatial data:
  - Find regions with high point intensity
  - Separated by areas with low intensity

Different approaches to clustering

- Four main approaches
- Partitioning
  - Task: divide the data into a given number of clusters
- Hierarchical
  - Create a tree based on the similarity / distance of items
- Density-based
  - Find contiguous areas with high density
- Grid-based
  - Divide the data space into grid cells

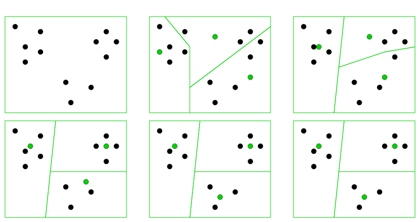
Clustering by partitioning

- Goal: divide the data into a pre-determined number of clusters
  - Several algorithms for this
- *k*-means
  - Presented in 1967
  - Start with *k* random points in the observation space
  - Repeat
    - Attach each observation to the closest of these points
    - Replace each of the *k* points with the centroid of the observations attached to it

    until the clustering stabilises

k-means: example

- Find three clusters



- After four iterations the clustering stabilises

Partitioning methods

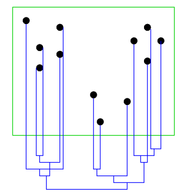
- *k*-medoids
  - Use the centermost cluster member instead of the mean
- EM (expectation maximisation)
  - Define the cluster by a probability distribution instead of a centroid
- Guaranteed to find a local optimum
- Not globally optimal clustering
  - Choice of the random seeds affects final result
- Tend to find circular clusters

Hierarchical clustering

- Form a tree of the data points
  - Each sub-tree contains points that are close to each other
  - These sub-trees can be considered as clusters
  - Resolution of clustering can be chosen afterwards
- Common alternatives
  - Agglomerative clustering: start from the bottom, join branches that are similar
  - Divisive clustering: start from the top, split branches that are dissimilar

Hierarchical clustering: example

- Agglomerative clustering
- Distance between clusters calculated as the average of distances between points

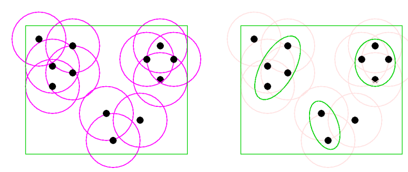


Density-based clustering

- Clustering no longer based on the distance between points
- Instead, density of points
- Good for finding non-spherical clusters
- Clusters do not necessarily cover all points

Density-based clustering

- A point in a high-density area belongs to a cluster
- In other words, a point belongs to a cluster if it is close enough to other points



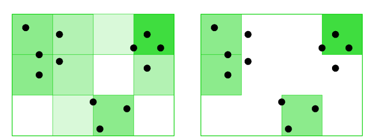
- More fine-grained density calculations may be useful

Grid-based clustering

- Goal: find areas with high point density
- Divide the space to grid cells
- Compute the density of points in each cell
- Use the high-density cells to define clusters
- More effective than density-based clustering

Grid-based clustering

- Calculate density in each grid cell
- Discard cells with too low density



All in all

- Several approaches to clustering

- Some are not really intended for spatial clustering
  - Original goal often finding patterns in
    multi-variable observation data
  - Nevertheless, possible to use for spatial data

- Challenges in spatial clustering
  - Arbitrarily shaped clusters
  - Euclidean distance not necessarily useful
    - Roads, water, other terrain effects

- Next week: a couple of methods for spatial data