NDA: Clustering

Maximilien Danisch

LIP6 – CNRS and Sobonne University

first_name.last_name@lip6.fr

Bibliography

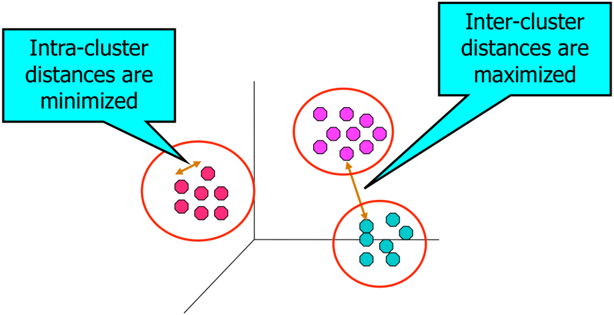
- Introduction to Data Mining, 2nd Edition by Tan, Steinbach, Karpatne, Kumar Chapter 7.

Outline

Outline

What is clustering?

Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



What is clustering?

- Clustering is an unsupervised learning method (i.e. no predefined classes)
- It is different from classification (supervised learning)

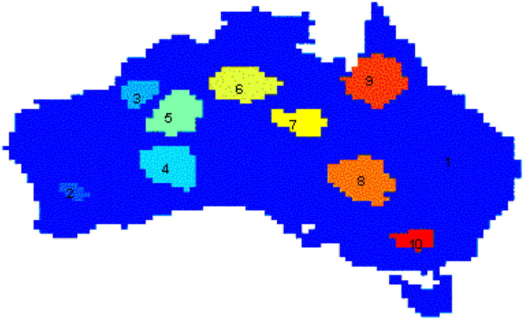
Why is it usefull (applications of cluster analysis)?

Understanding the data / get insights on the data:

- Group related documents for browsing
- Group genes and proteins that have similar functionality
- Group people sharing similar interest
- Group movies with similar genres or actors

Why is it usefull (applications of cluster analysis)?

Summarization: Reduce the size of large data sets



Clustering precipitation in Australia

Custering is an ill-defined problem

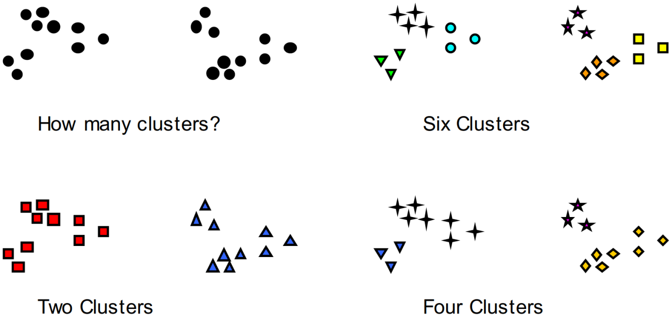
How many clusters?

How many clusters?

- An Impossibility Theorem for Clustering, Jon Kleinberg, NIPS2003.

Custering is an ill-defined problem

How many clusters?



- An Impossibility Theorem for Clustering, Jon Kleinberg, NIPS2003.

Several types of clusterings

A clustering is a set of clusters. Important distinction between hierarchical and partitional sets of clusters.

- **Partitional clustering:** A division of data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset
- **Hierarchical clustering:** A set of nested clusters organized as a hierarchical tree

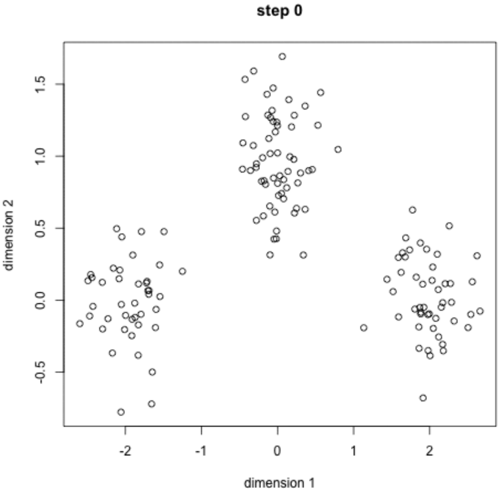
Several types of clusterings

Other distinctions:

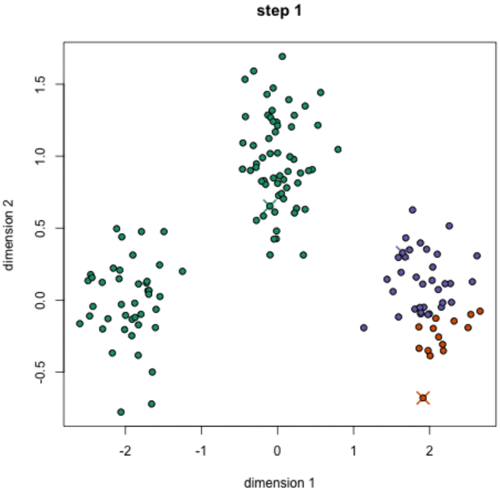
- **Exclusive versus non-exclusive:** In non-exclusive clusterings, points may belong to multiple clusters. Can represent multiple classes or ‘border’ points
- **Fuzzy versus non-fuzzy:** In fuzzy clustering, a point belongs to every cluster with some weight between 0 and 1. Weights must sum to 1
- **Partial versus complete:** In some cases, we only want to cluster some of the data
- **Heterogeneous versus homogeneous:** Clusters of widely different sizes, shapes, and densities

Outline

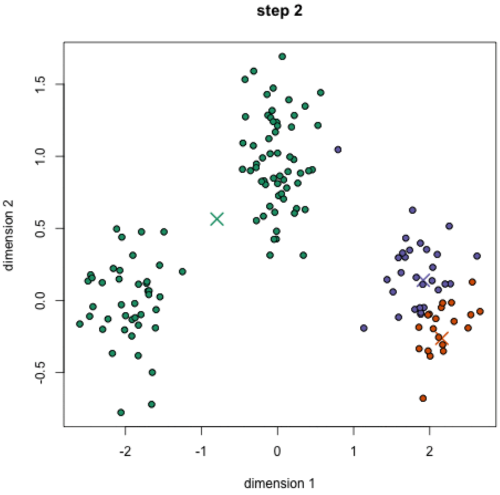
k-means



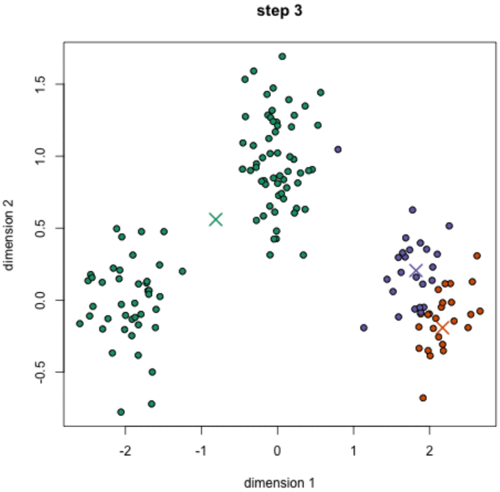
k-means



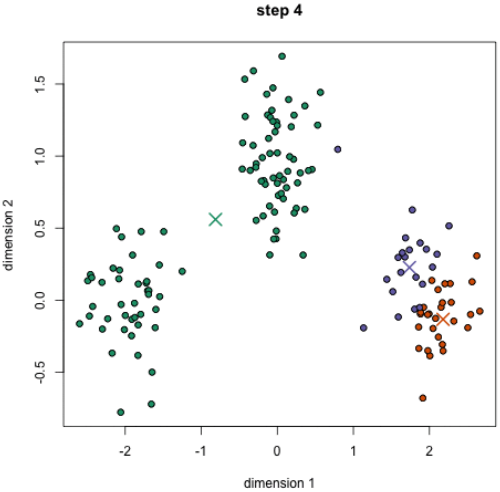
k-means



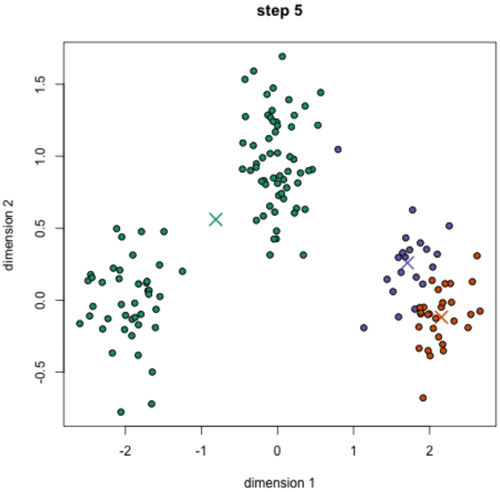
k-means



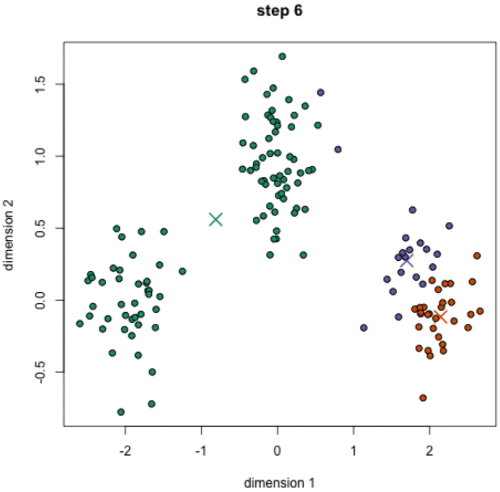
k-means



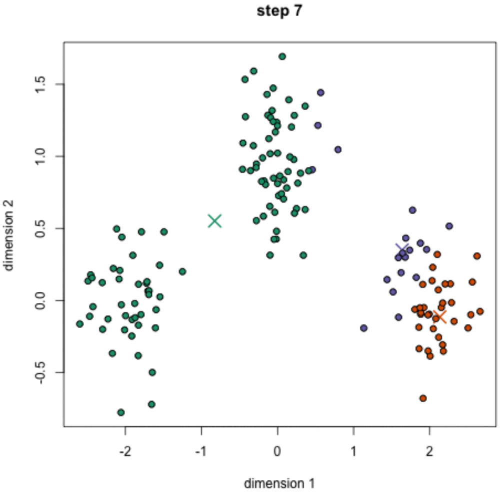
k-means



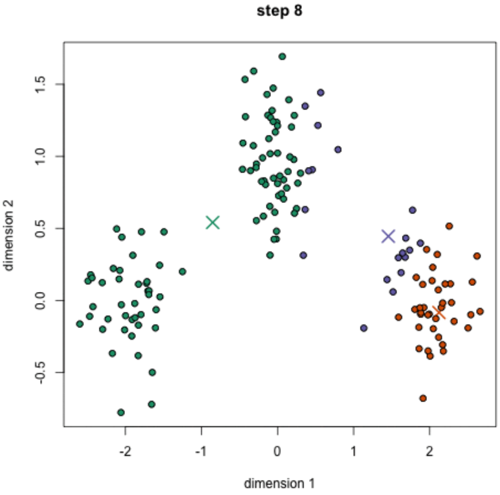
k-means



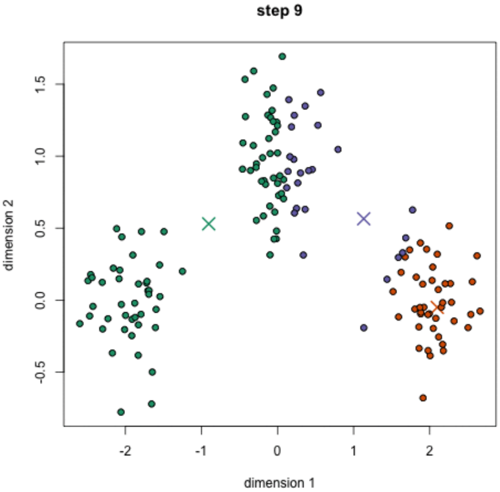
k-means



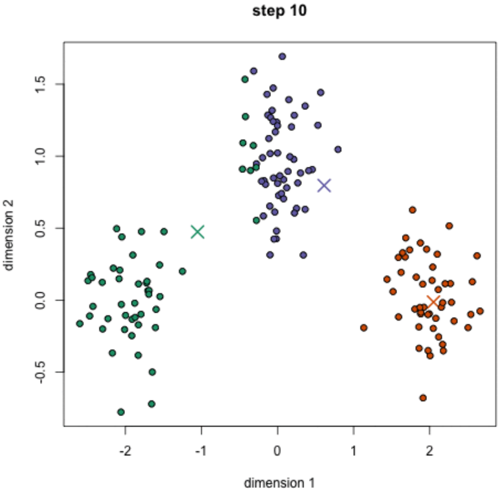
k-means



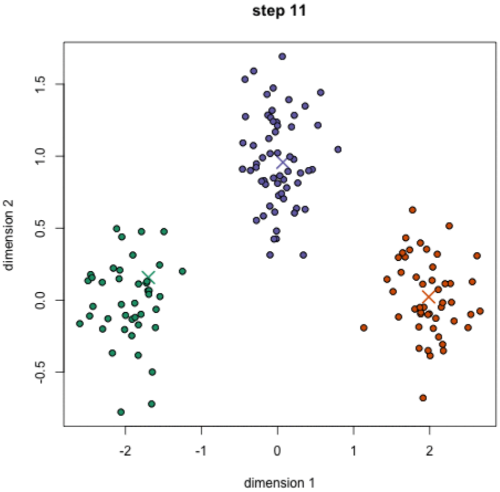
k-means



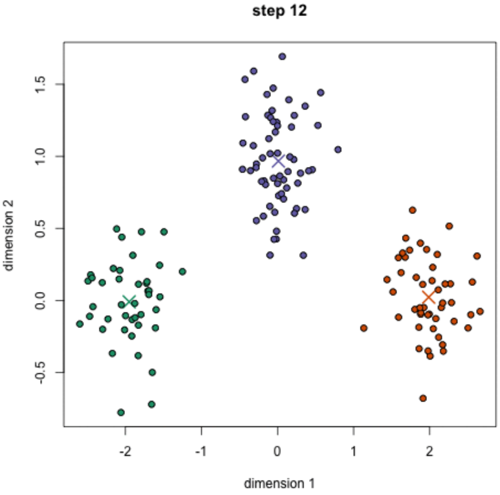
k-means



k-means



k-means

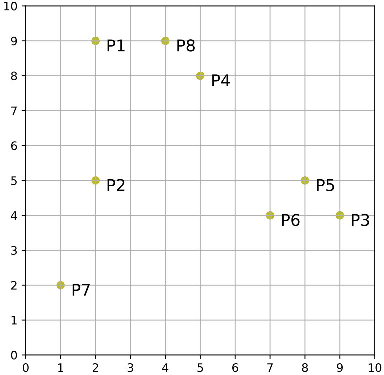


k-means

- Randomly chose k initial centroids
- While True:
  - Create k clusters by assigning each point to closest centroid
  - Compute k new centroids by averaging points in each clustering
  - If centroids don't change:
    - Break

k-means

Exercice: simulate some executions of kmeans wth $k = 3$ on the following set of points



k-means: distortion (a.k.a. SSE or SSD)

k-means can be seen as a heuristic to minimize the distortion:

$$\text{distortion} = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{i,j} ||x_i - \mu_j||_2^2$$

with

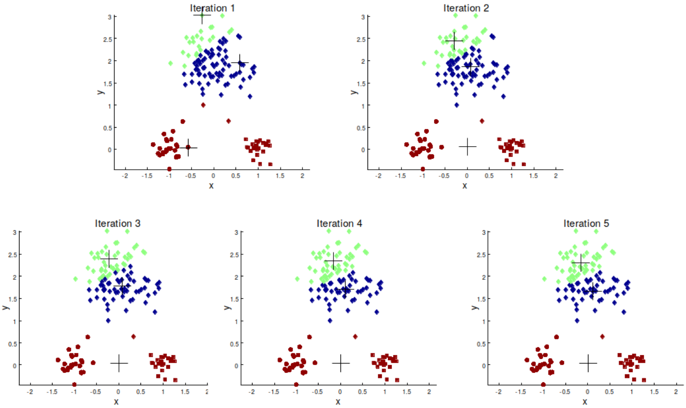
- μ_j the vector of centroid j and
- $w_{i,j} = 1$ if the sample x_i is in cluster j and 0 otherwise.

Limitation of k-means: clusters

- K-means has problems when clusters are of differing
  - Sizes
  - Densities
  - Non-globular shapes
- K-means has problems when the data contains outliers.

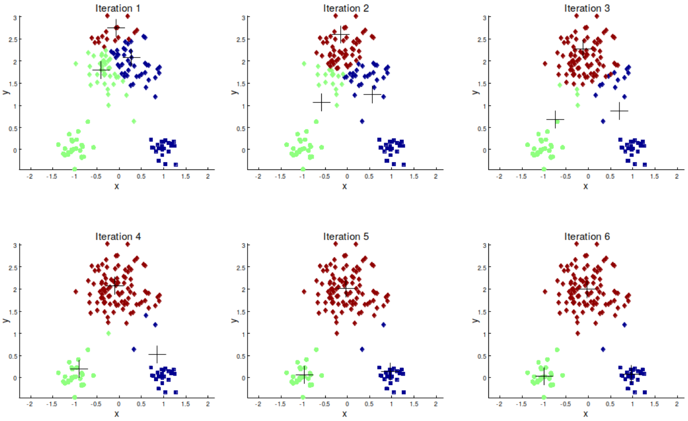
Normalising the data and removing outliers can help!

Limitation of k-means: initialisation



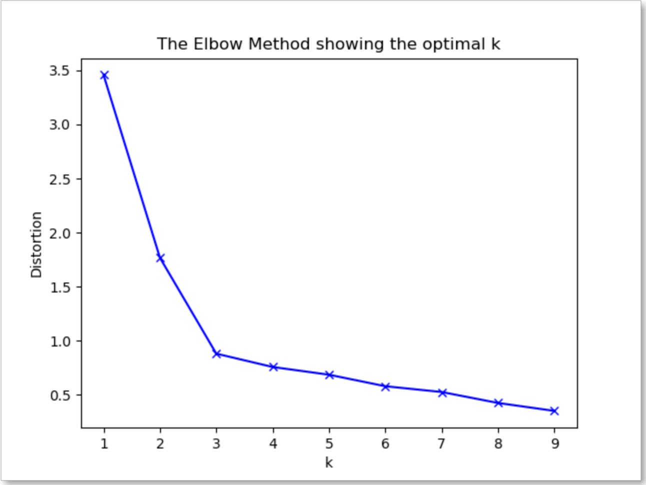
Using multiple runs or kmeans++ can help!

Limitation of k-means: initialisation



Using multiple runs or kmeans++ can help!

Limitation of k-means: How to choose k?



The elbow method: choose k=3, where the elbow is located

Limitation of k-means: How to choose k?

Silhouette value. A measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).

For data point i in cluster C_k, let

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \in C_k, i \neq j} d(i,j) \quad \text{and} \quad b(i) = \min_{l \neq k} \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j)$$

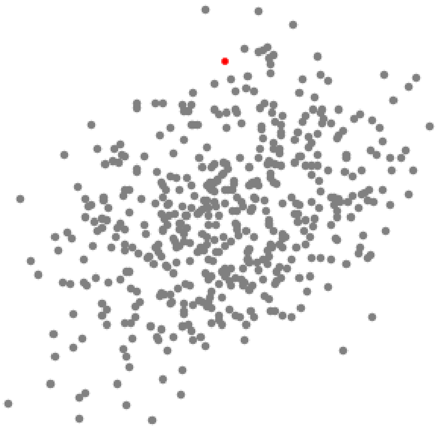
The silhouette score of one data point i: $s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$

Silouette score of a partition = average of the $s(i)$'s.

- **Elbow method with silhouette score instead of distortion**

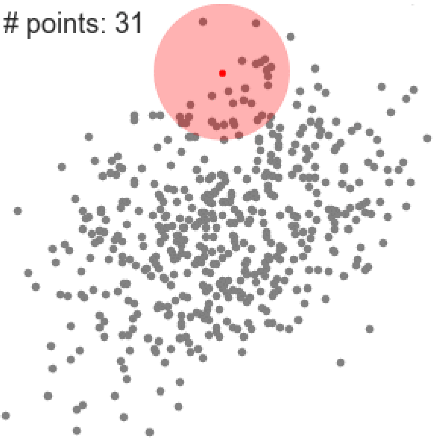
Outline

Mean-shift (with one centroid)



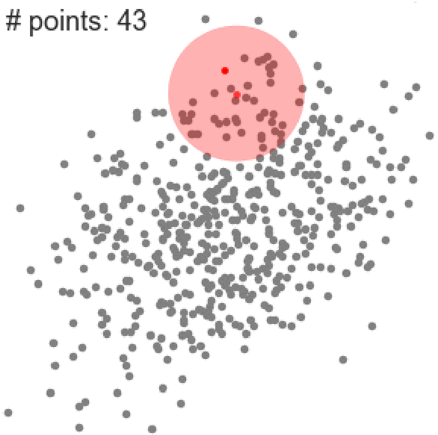
The centroid moves towards a higher density region

Mean-shift (with one centroid)



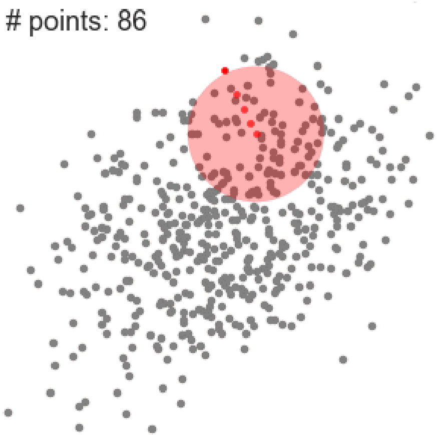
The centroid moves towards a higher density region

Mean-shift (with one centroid)



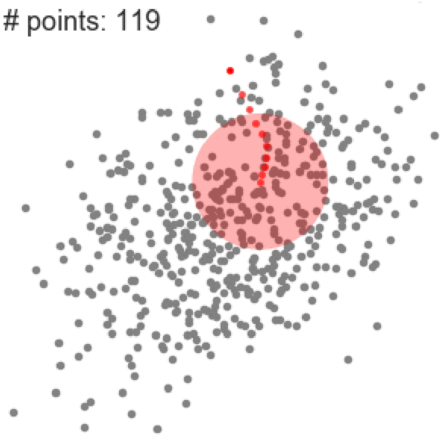
The centroid moves towards a higher density region

Mean-shift (with one centroid)



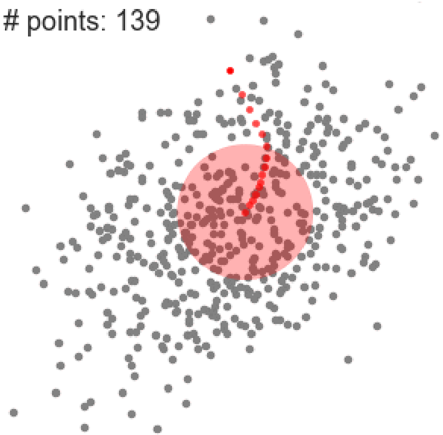
The centroid moves towards a higher density region

Mean-shift (with one centroid)



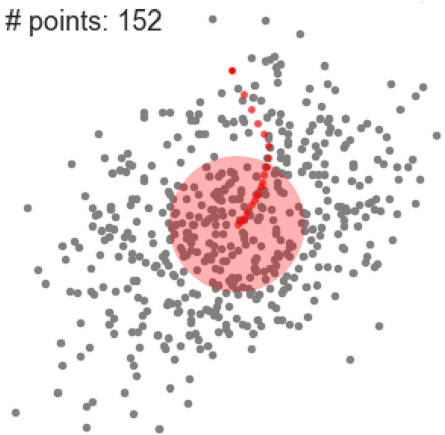
The centroid moves towards a higher density region

Mean-shift (with one centroid)



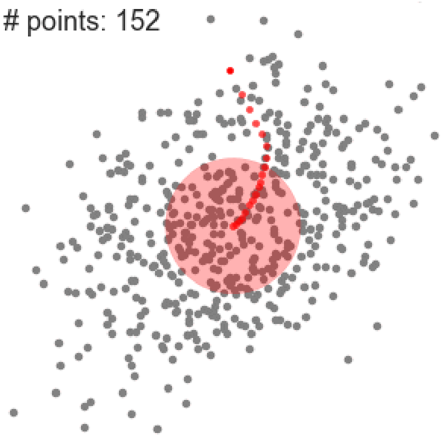
The centroid moves towards a higher density region

Mean-shift (with one centroid)



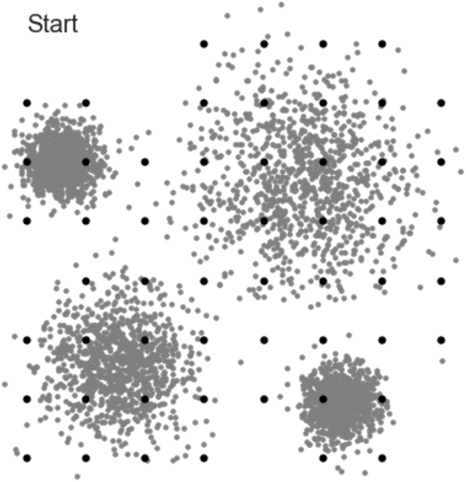
The centroid moves towards a higher density region

Mean-shift (with one centroid)

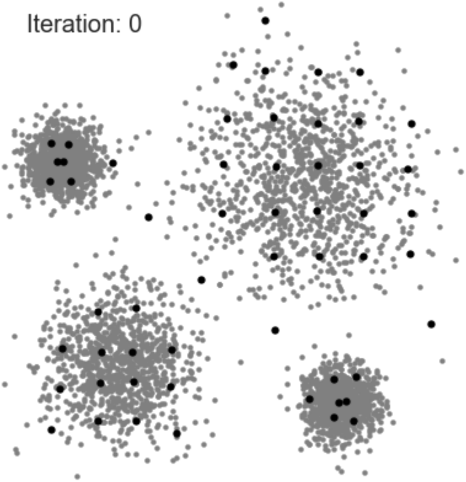


The centroid moves towards a higher density region

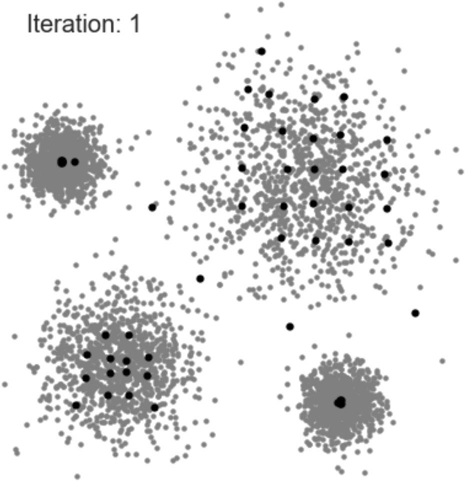
Mean-shift



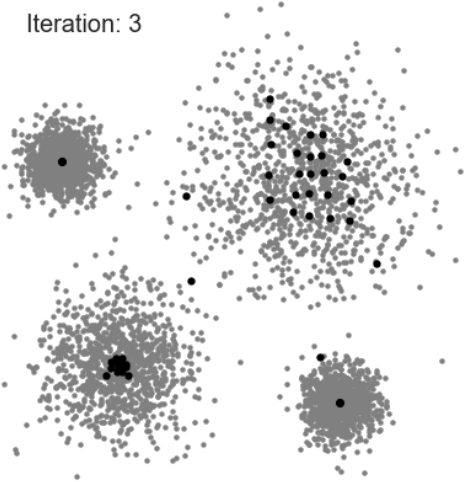
Mean-shift



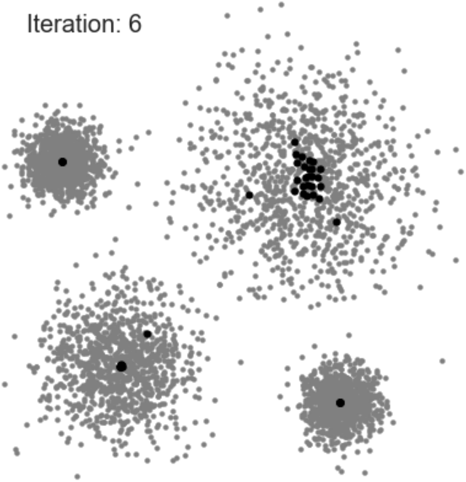
Mean-shift



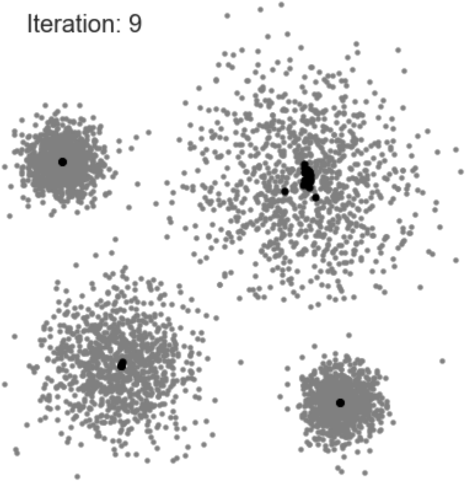
Mean-shift



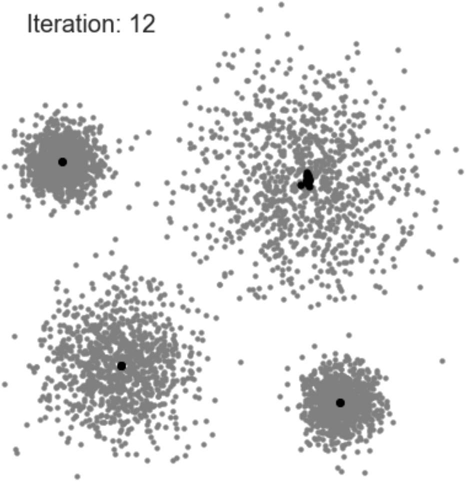
Mean-shift



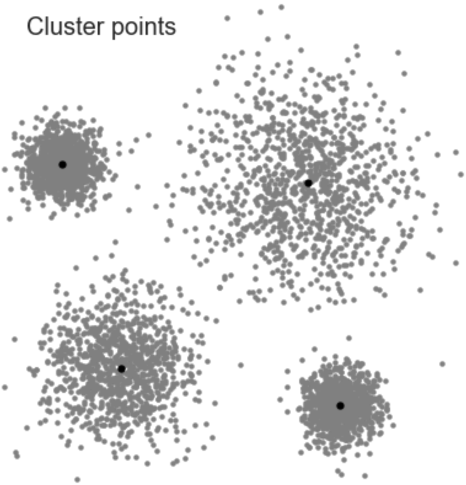
Mean-shift



Mean-shift



Mean-shift

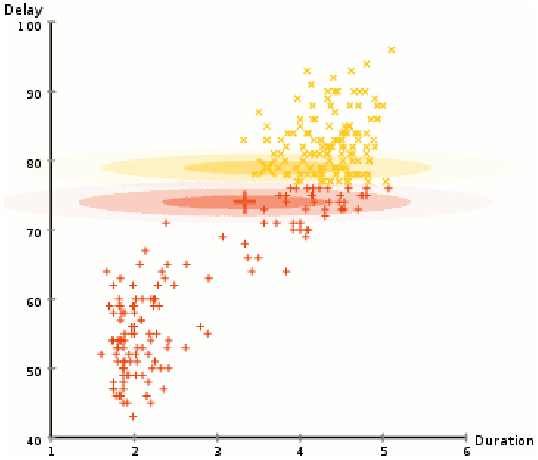


Mean-shift

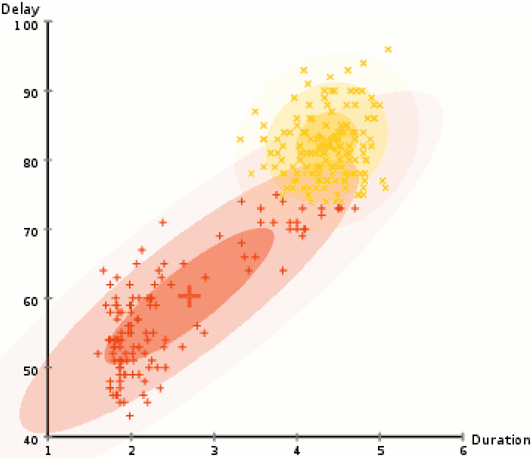
Mean-shift

- Start with a given (large) number of circular sliding windows centered at randomly selected centroids and having radius *r*.
- While True;
  - Compute k new centroids by averaging examples in each sliding windows (the centroids are shifted towards regions of higher density)
  - If centroids don't change:
    - Break
- If multiple sliding windows overlap, then only the window containing the most points is preserved.
- Each data point is assigned to the nearest centroid.

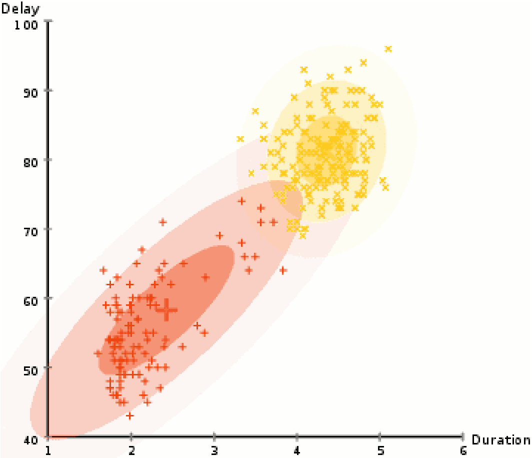
GMM: Gaussian Mixture Model



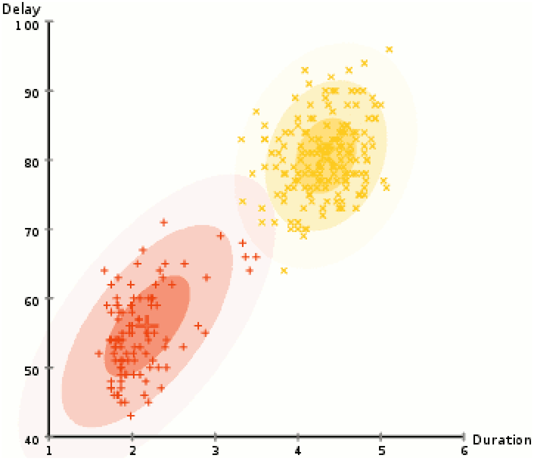
GMM: Gaussian Mixture Model



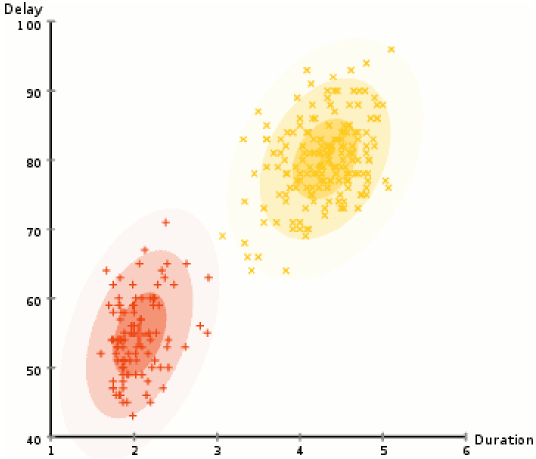
GMM: Gaussian Mixture Model



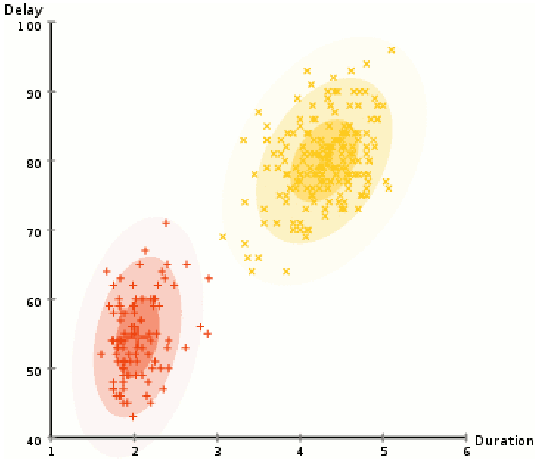
GMM: Gaussian Mixture Model



GMM: Gaussian Mixture Model



GMM: Gaussian Mixture Model



GMM: Gaussian Mixture Model

- Model: K gaussians (one-dimentional):
  - $p(x) = \sum_{k=1}^{K} \Phi_k N(x|\mu_k, \sigma_k)$
  - $N(x|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$
  - $\sum_{k=1}^{K} \Phi_k = 1$
- Want to find $(\Phi_k, \mu_k, \sigma_k)$ for all k maximizing $\prod_{i=1}^{n} p(x_i)$
- Chicken and egg problem:
  - need $(\Phi_k, \mu_k, \sigma_k)$ for all k to guess source of points
  - need to know source to estimate $(\Phi_k, \mu_k, \sigma_k)$

GMM: Gaussian Mixture Model

Expectation-Maximization algorithm:

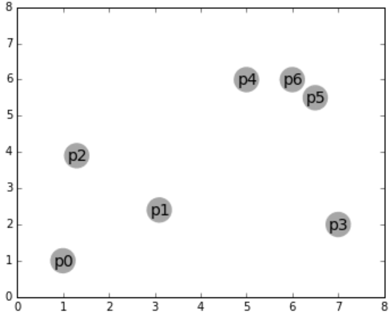
- start with randomly placed Gaussians $(\Phi_k, \mu_k, \sigma_k)$
- (E-step) for each i, k compute $\gamma_{i,k} \sim$ did x_i came from k?
- (M-step) adjust $(\Phi_k, \mu_k, \sigma_k)$ to fit points assigned to them
- iterate until convergence

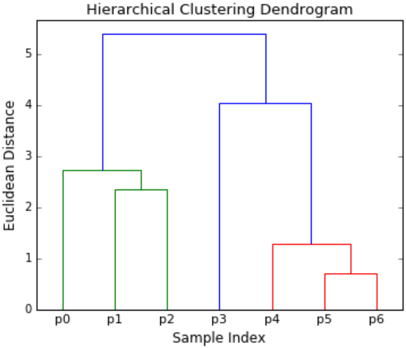
$$\Phi_k = \sum_{i=1}^{N} \frac{\gamma_{i,k}}{N}$$

$$\gamma_{i,k} = \frac{\Phi_k \mathcal{N}(x_i|\mu_k,\sigma_k)}{\sum_{k=1}^{K} \Phi_k(x|\mu_k,\sigma_k)}$$

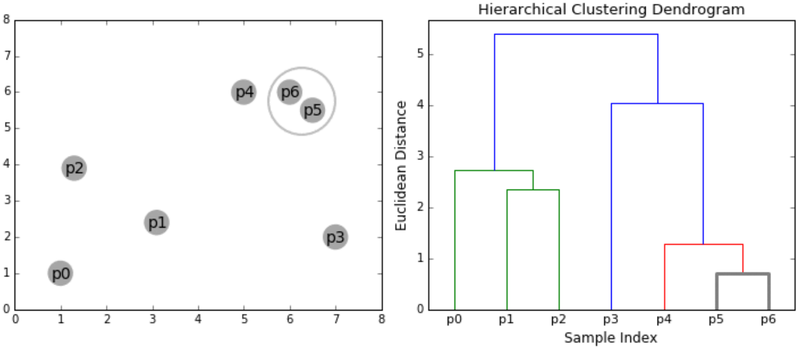
$$\mu_k = \frac{\sum_{i=1}^{N} \gamma_{i,k} x_i}{\sum_{i=1}^{N} \gamma_{i,k}} \qquad \sigma_k^2 = \frac{\sum_{i=1}^{N} \gamma_{i,k}(x_i-\mu_k)^2}{\sum_{i=1}^{N} \gamma_{i,k}}$$

Hierarchical clustering

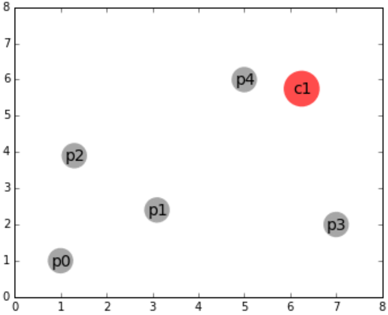


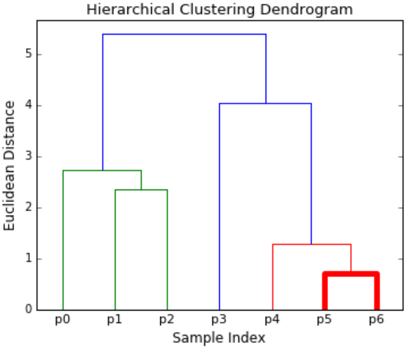


Hierarchical clustering

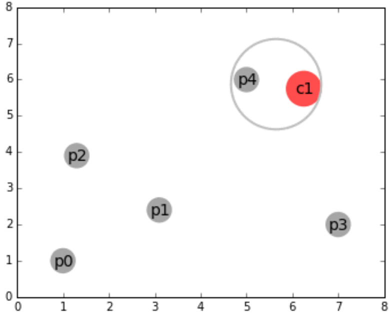


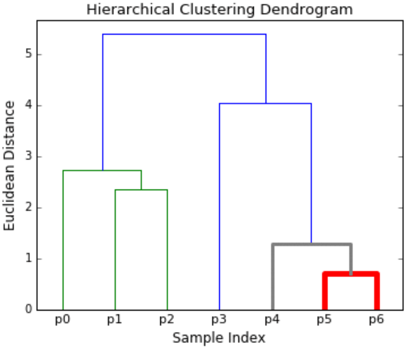
Hierarchical clustering



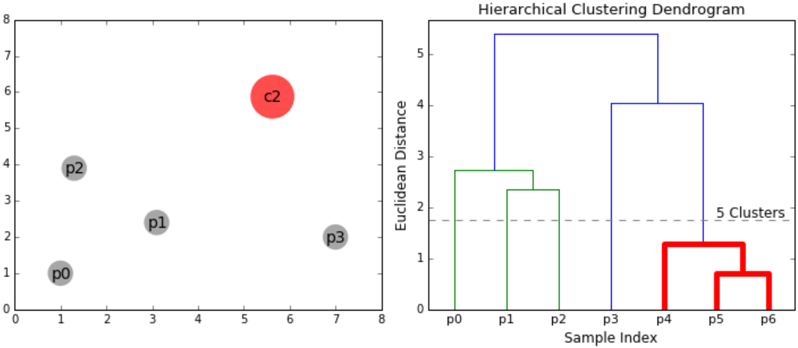


Hierarchical clustering

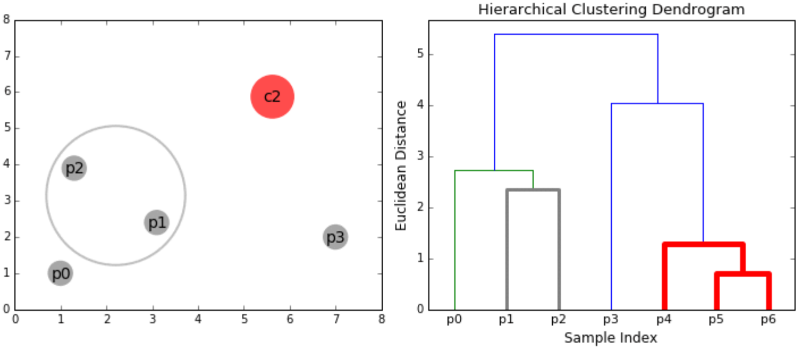




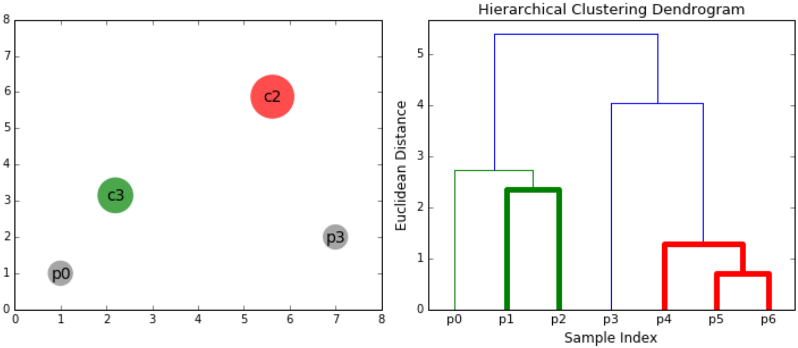
Hierarchical clustering



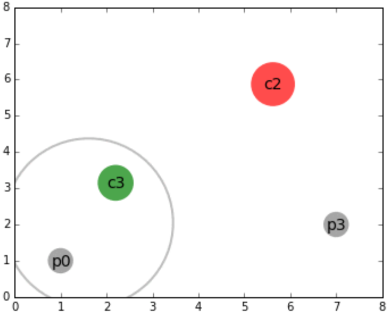
Hierarchical clustering

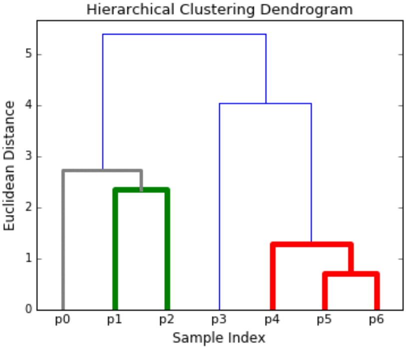


Hierarchical clustering

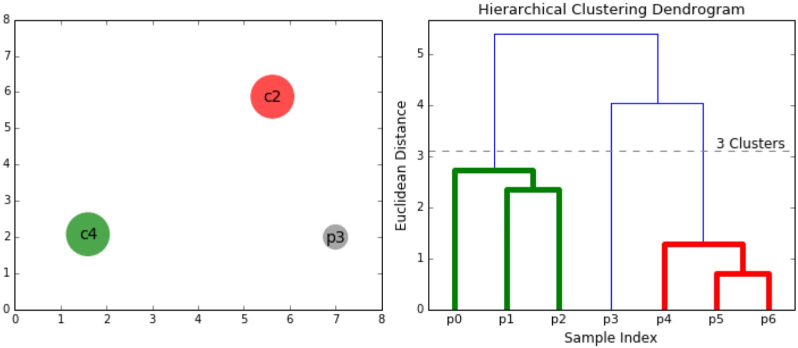


Hierarchical clustering

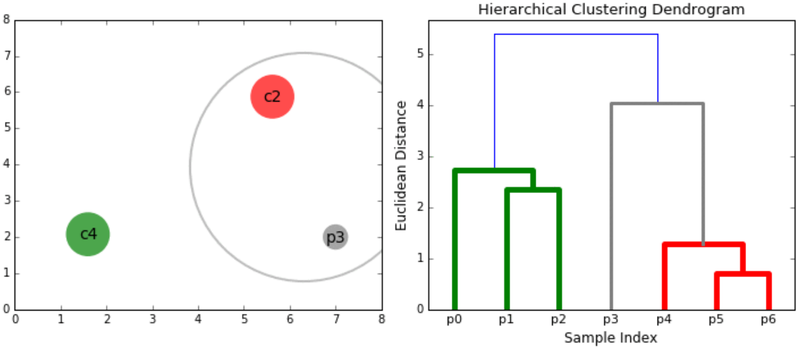




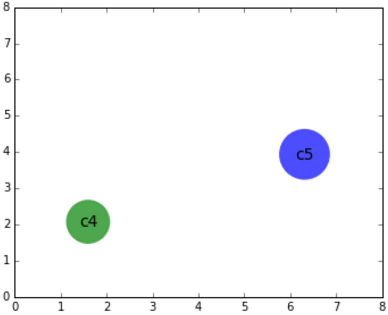
Hierarchical clustering

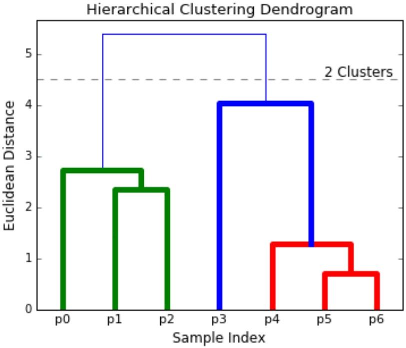


Hierarchical clustering

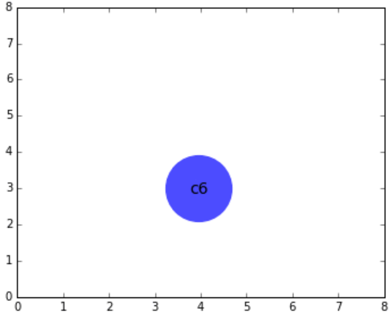


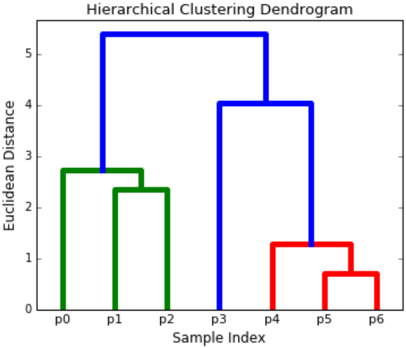
Hierarchical clustering





Hierarchical clustering





Hierarchical clustering

Two types of hierarchical clustering:

1. **Agglomerative.** This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
2. **Divisive.** This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Aglomerative hierarchical clustering

- Maximum (or complete) linkage clustering:

$$d(A, B) = \max\{d(a, b) : a \in A, b \in B\}$$

- Minimum (or single) linkage clustering:

$$d(A, B) = \min\{d(a, b) : a \in A, b \in B\}$$

- Average linkage clustering:

$$d(A, B) = \frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$$

Divisive hierarchical clustering

Dasgupta's objective and algorithm: Repeatedly subdivides the elements using an approximation algorithm for the sparsest cut problem.

Outline

1. What is clustering? Motivation?
2. k-means
3. Other clustering algorithms
   - Mean-shift
   - GMM: Gaussian Mixture Model
   - Hierarchical clustering
4. Validation

Validation

- For supervised classification we have a variety of measures to evaluate how good our model is: Accuracy, precision, recall
- For clustering, the analogous question is how to evaluate the “goodness” of the resulting clusters?
- "Clusters are in the eye of the beholder"!
- We still want some tools to:
  - To avoid finding patterns in noise
  - To compare clustering algorithms
  - To compare two sets of clusters
  - To compare two clusters

Validation

Two types of numerical measures to judge cluster validity:

1. Internal Index: Used to measure the goodness of a clustering structure without respect to external information. (e.g. distortion, silhoutte score)
2. External Index: Used to measure the extent to which cluster labels match externally supplied class labels. (e.g. Entropy, Purity, Adjusted Rand Index)

Validation: External Index

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

m_j = size of cluster j, m = number of documents
p_{ij} = probability that a random document of cluster j belongs to topic i. For example, $p_{13} = 1/685$.

- Entropy of a cluster:

$$e_j = -\sum_{i=1}^{n} p_{ij} \log_2(p_{ij})$$

- Entropy of a clustering:

$$\sum_j \frac{m_j}{m} e_j$$

- Purity of a cluster:

$$purity_j = max_i(p_{ij})$$

- Purity of a clustering:

$$\sum_j \frac{m_j}{m} purity_j$$