CS4801 : Unsupervised methods

Sahely Bhadra

25/9/2018

1. Introduction to Clustering
2. Different type of clustering methods
3. k-means clustering

SELF READING : SVMs for regression (7.1.4 in Bishop's)

Unsupervised learning

- Machine learning algorithm used to **draw inferences** from datasets consisting of **input data** [x_1,…x_n] **without labeled** responses.

- Goal : The goal is to discover interesting things about the measurements/ data :

    - Is there an informative way to visualise the data?

    - Can we discover subgroups among the features or among the observations/ samples?

- Example :

    - **Cluster analysis** : a broad class of methods for discovering unknown subgroups in data

    - **Principal Component Analysis** : a tool used for data visualization or data pre-processing before supervised techniques are applied : Gene selection, data visualisation

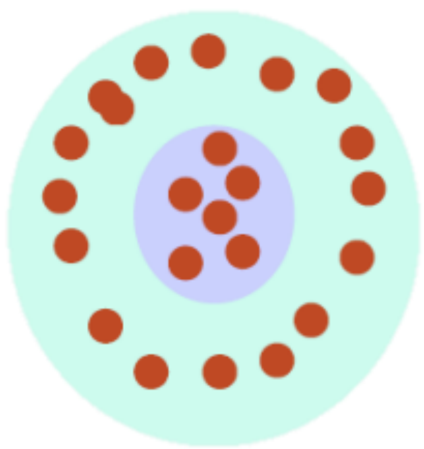
Challenges in unsupervised learning

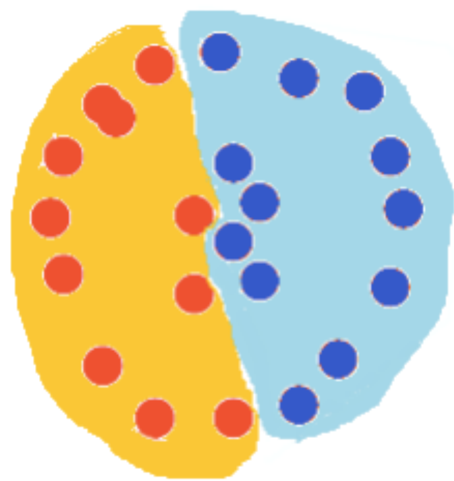
- Unsupervised learning is more subjective than supervised learning, as there is no simple goal for the analysis, such as prediction of a response.
- But techniques for unsupervised learning are of growing importance in a number of fields:
- Examples :
    - Image segmentation : segment different parts of an image
    - subgroups of breast cancer patients grouped by their gene expression measurements
    - groups of shoppers characterized by their browsing and purchase histories,
    - movies grouped by the ratings assigned by movie viewers.

What is clustering

- The organisation of **unlabelled** data into **similarity** groups called clusters.

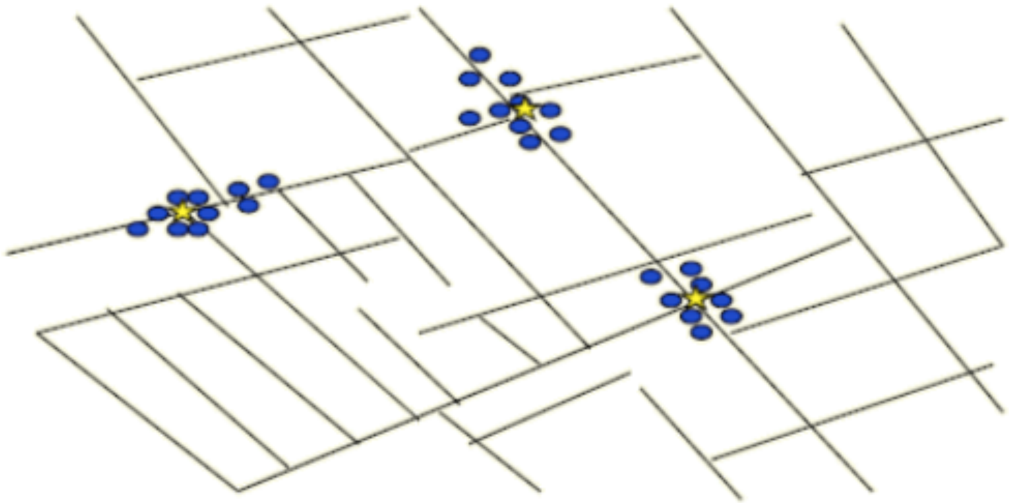
- A cluster is a collection of data items which are **"similar" between them**, and **"dissimilar"** to data items in **other clusters**.





History of clustering

- John Snow, a London physician plotted the location of cholera deaths on a map during an outbreak in the 1850s.
- The locations indicated that cases were clustered around certain intersections where there were polluted wells -- thus exposing both the problem and the solution.



From: Nina Mishra HP Labs

Clustering example

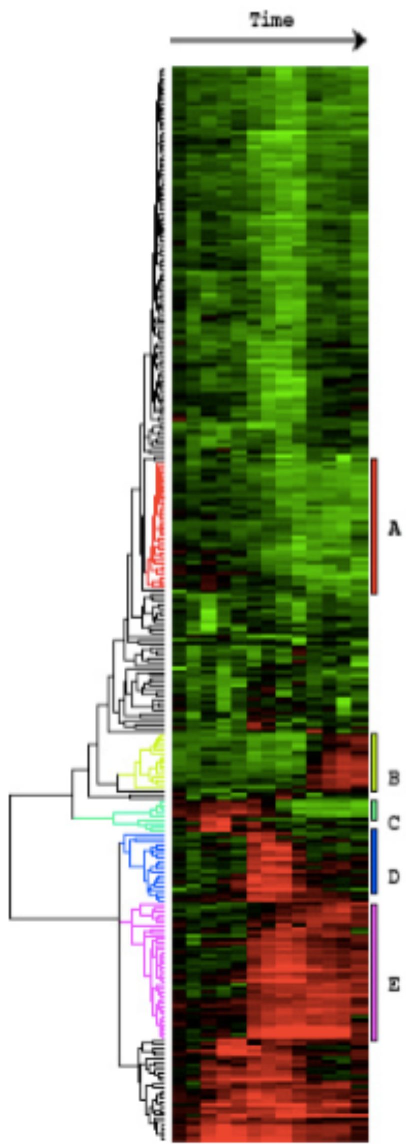
Image segmentation

Goal: Break up the image into meaningful or perceptually similar regions

[Slide from James Hayes]

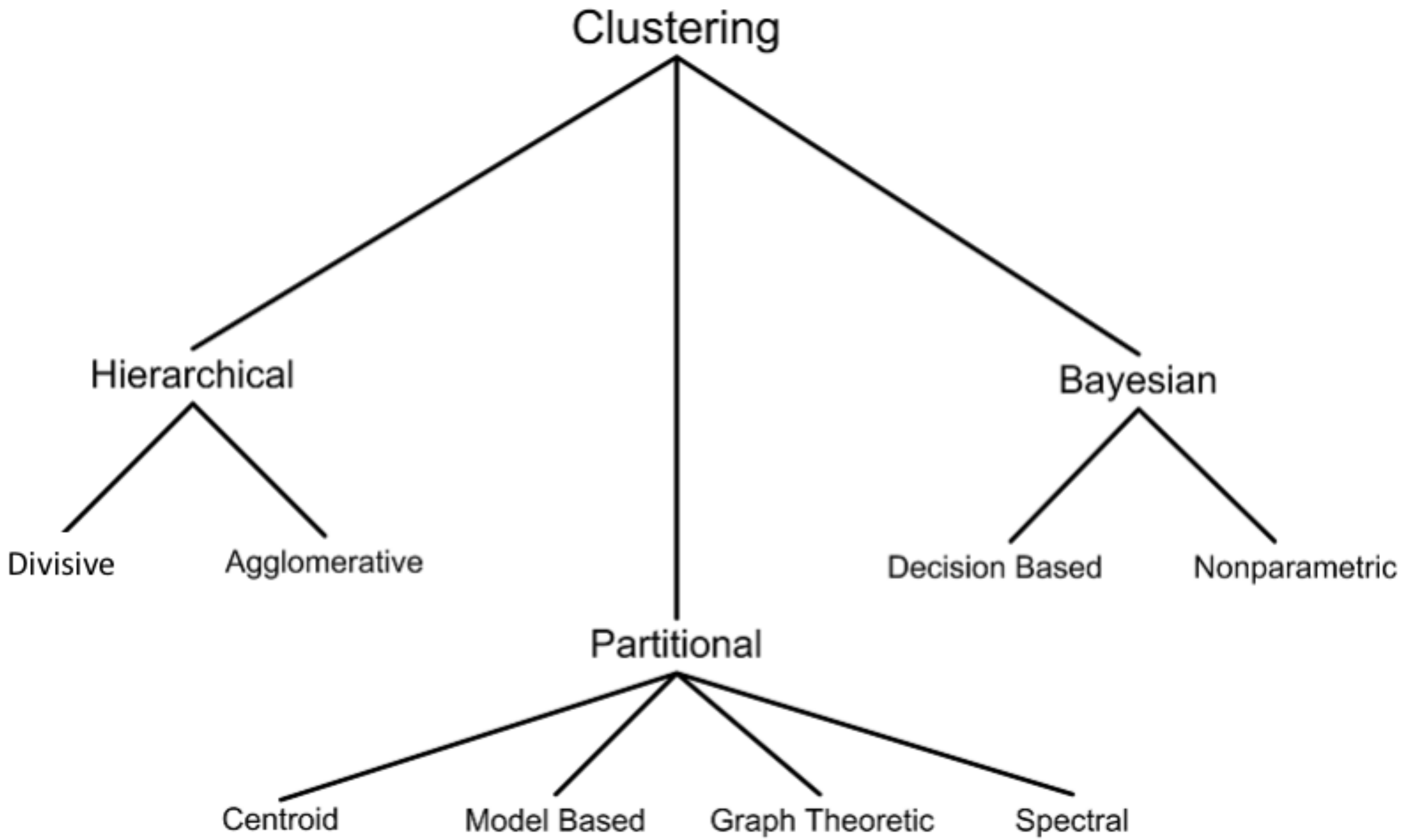
Clustering example

Clustering gene expression data



Eisen et al, PNAS 1998

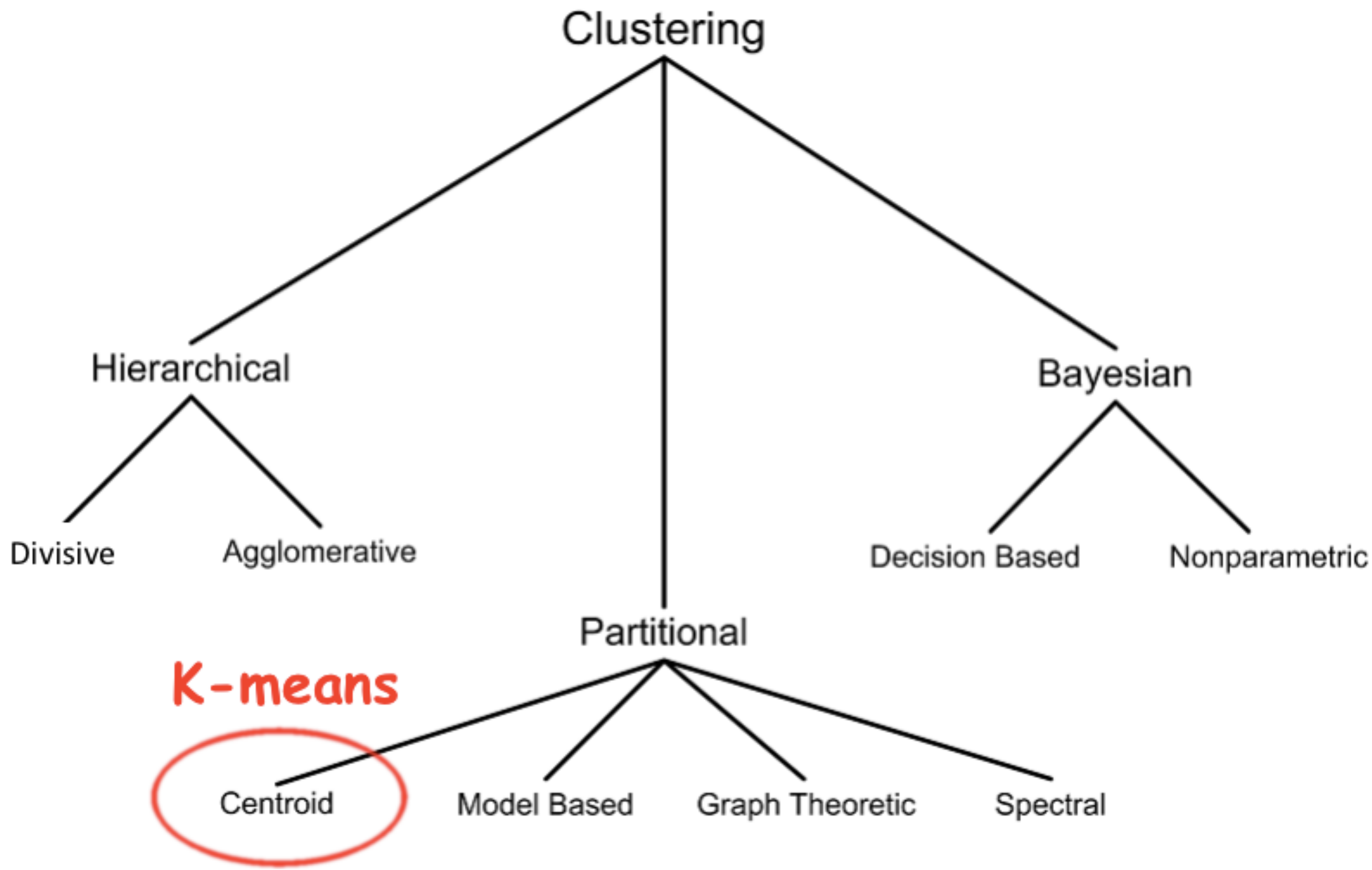
Clustering techniques



Clustering techniques

- **Hierarchical** algorithms find successive clusters using previously established clusters. These algorithms can be either agglomerative ("*bottom-up*") or divisive ("*top-down*"):
    1. Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters;
    2. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters.
- **Partitional** algorithms typically determine all clusters at once
- **Bayesian** algorithms try to generate a *posteriori distribution* over the collection of all partitions of the data.

Clustering techniques



K-means Clustering

K-means (MacQueen, 1967) is a partitional clustering algorithm

Let the set of data points D be $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$,

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ is a vector in $X \subseteq R^p$, and p is the number of dimensions.

The k-means algorithm partitions the given data into k clusters:

- Each cluster has a cluster **center**, called **centroid**.
- k is specified by the user

Gaol : Minimising Within cluster variance

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}.$$

K-means Clustering : Algorithm

Given k, the *k-means* algorithm works as follows:

1. Choose k (random) data points (seeds) to be the initial centroids, cluster centers
2. Assign each data point to the closest centroid
3. Re-compute the centroids using the current cluster memberships
4. If a convergence criterion is not met, repeat steps 2 and 3

K-means Clustering : When to stop

no (or minimum) re-assignments of data points to different clusters, *or*

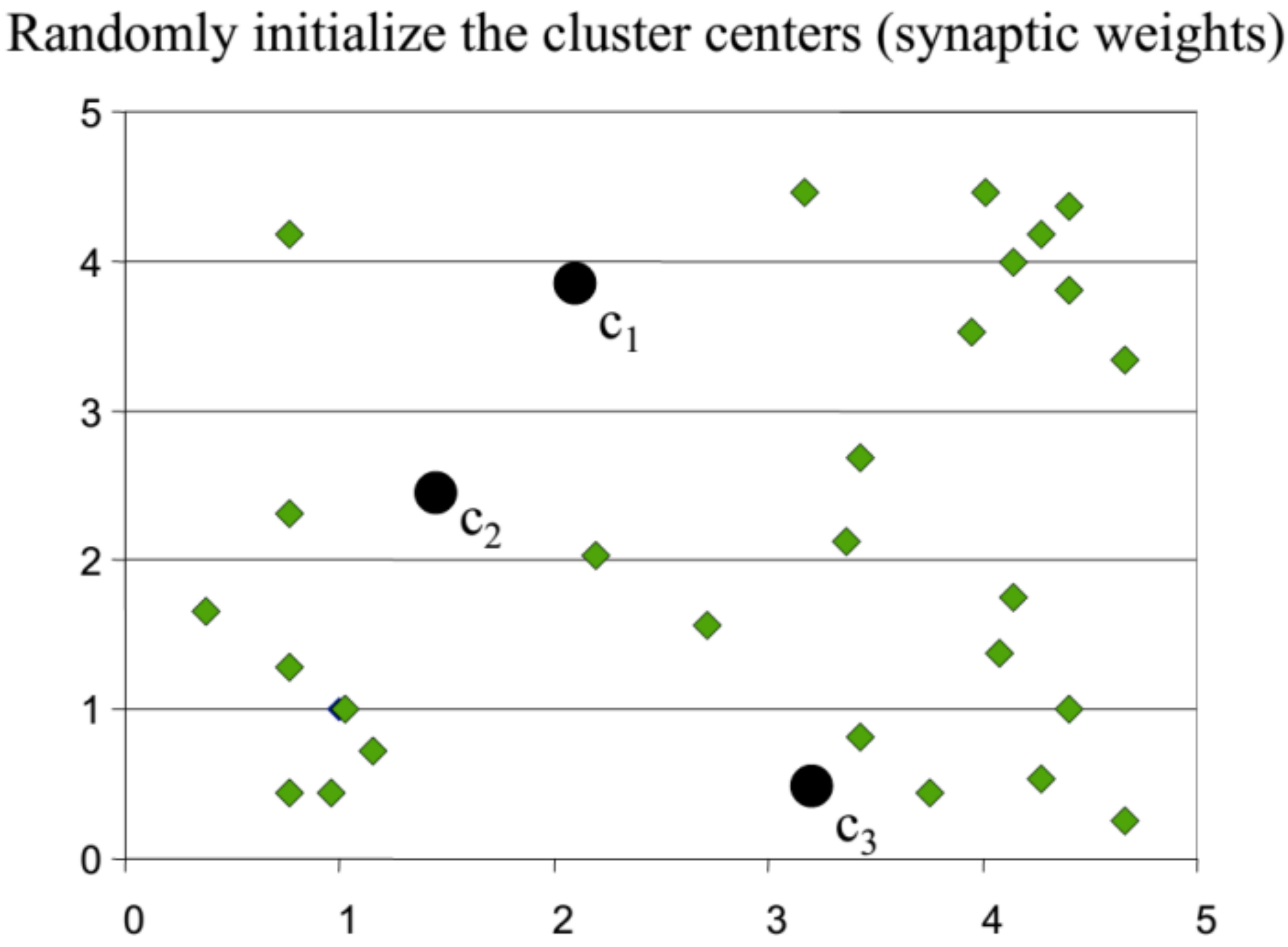
no (or minimum) change of centroids, *or*

minimum decrease in the **sum of squared error** (SSE),

$$SSE = \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2$$

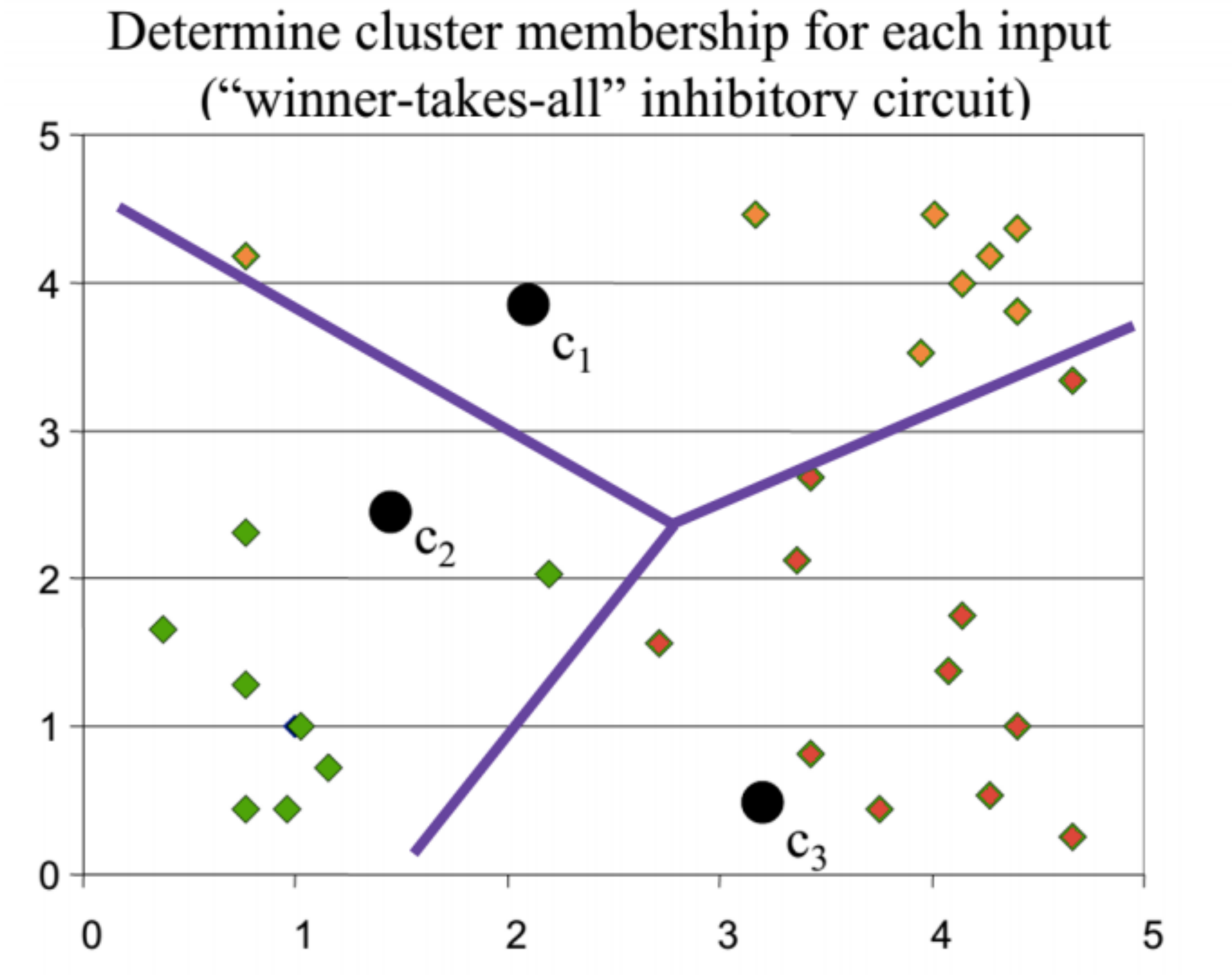
where $\bar{x}_{kj} = \frac{1}{|C_k|} \sum_{i \in C_k} x_{ij}$ is the mean for feature j in cluster C_k.

K-means Clustering : Example



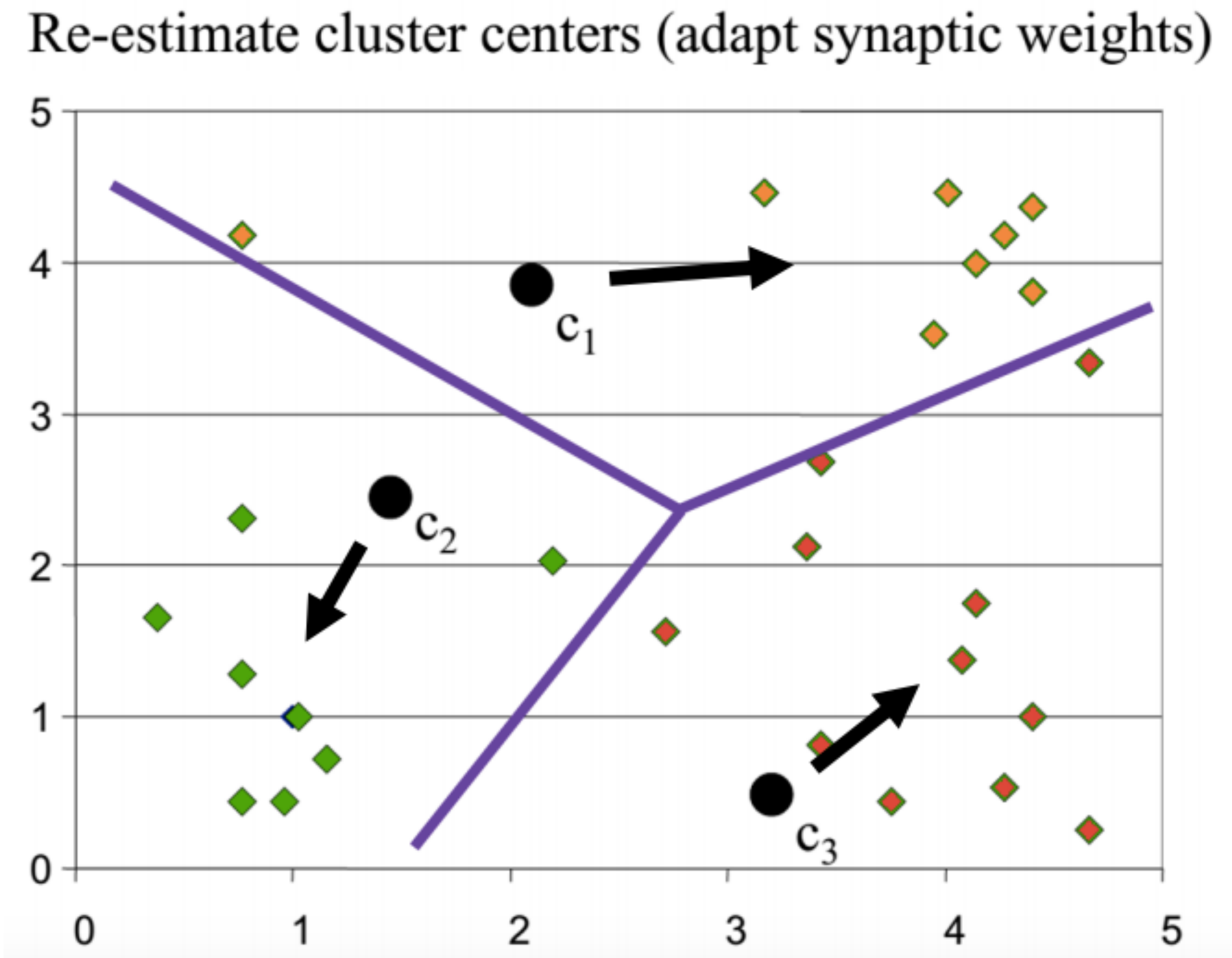
Step1

K-means Clustering : Example



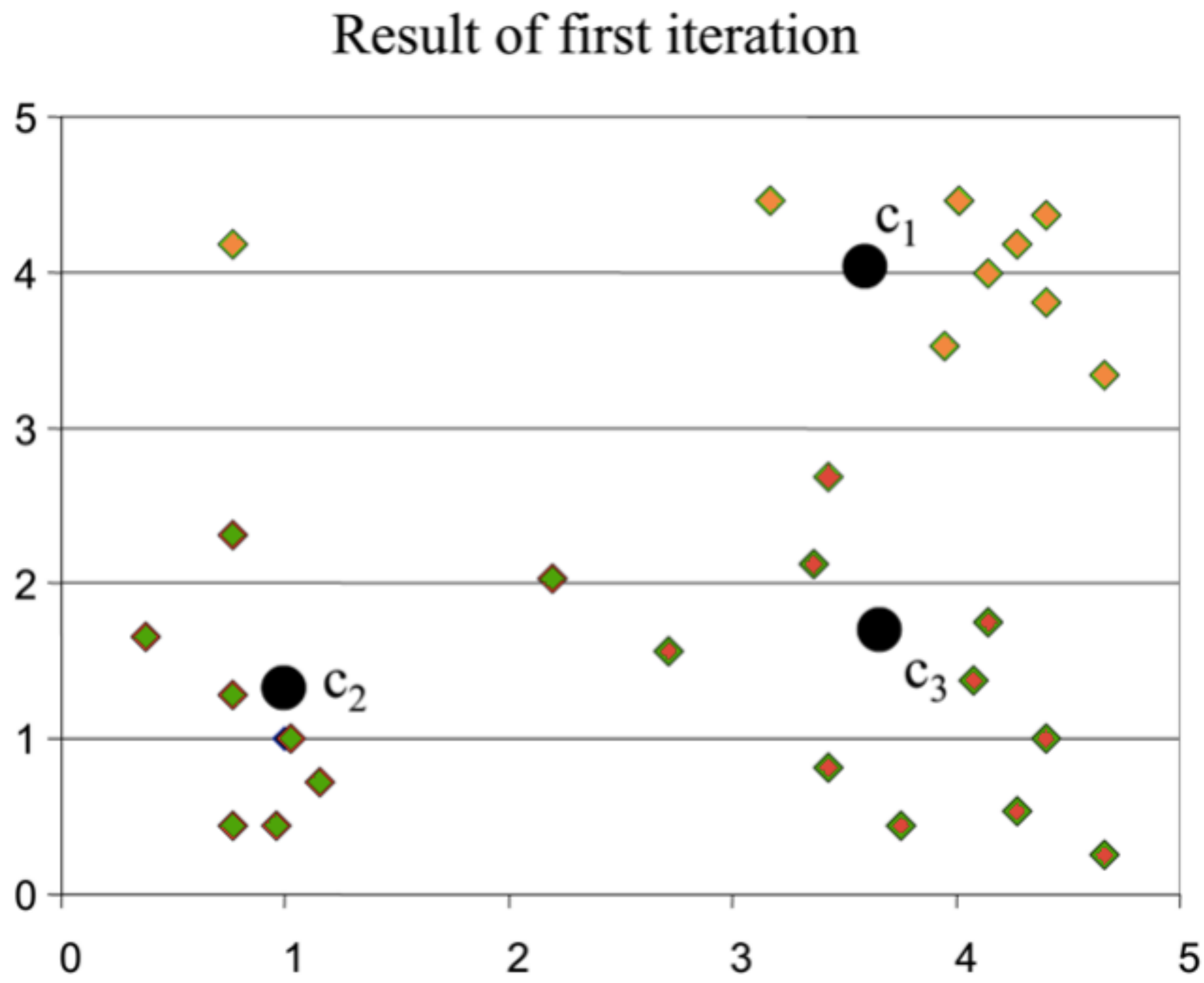
Step2

K-means Clustering : Example

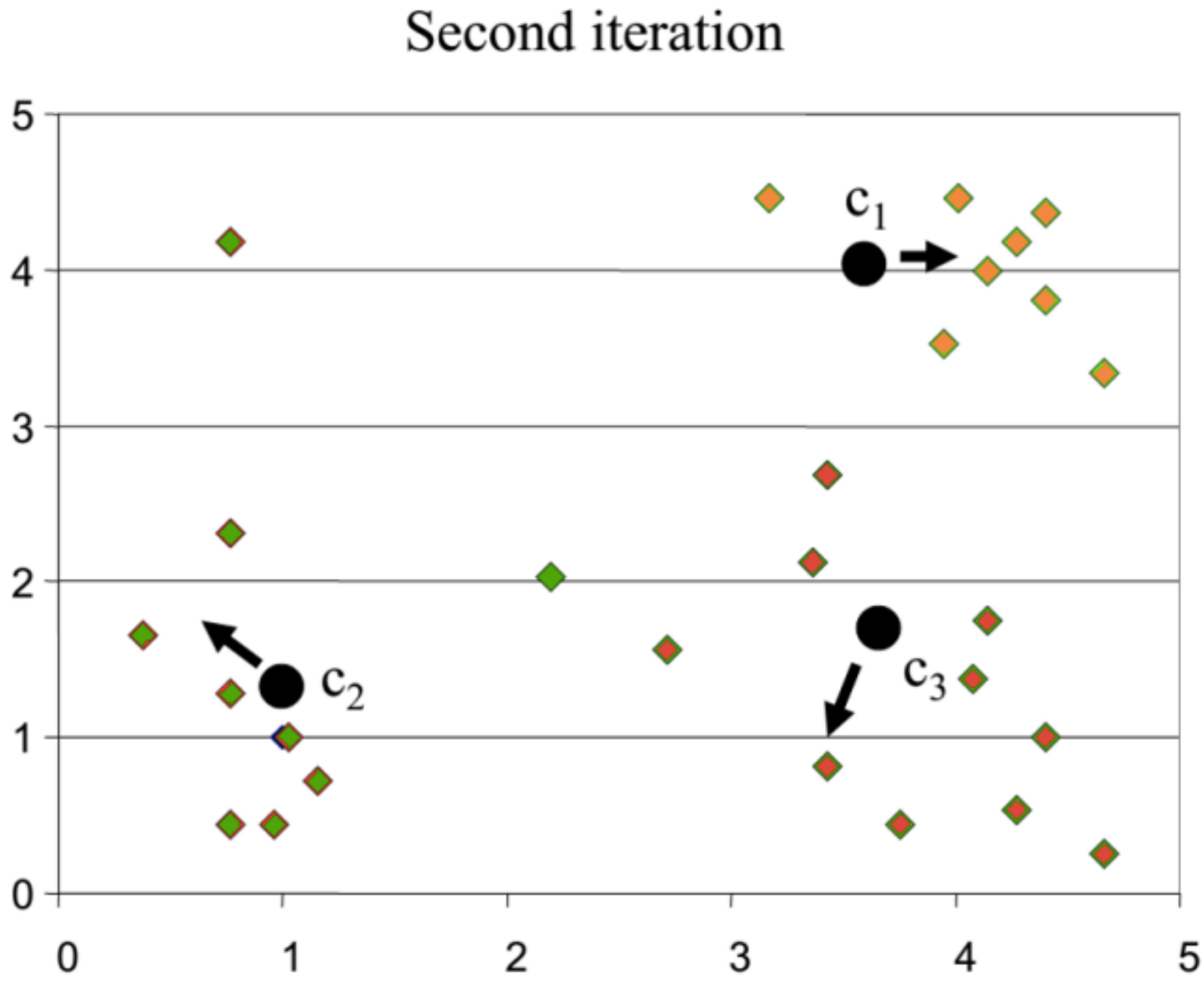


Step3

K-means Clustering : Example

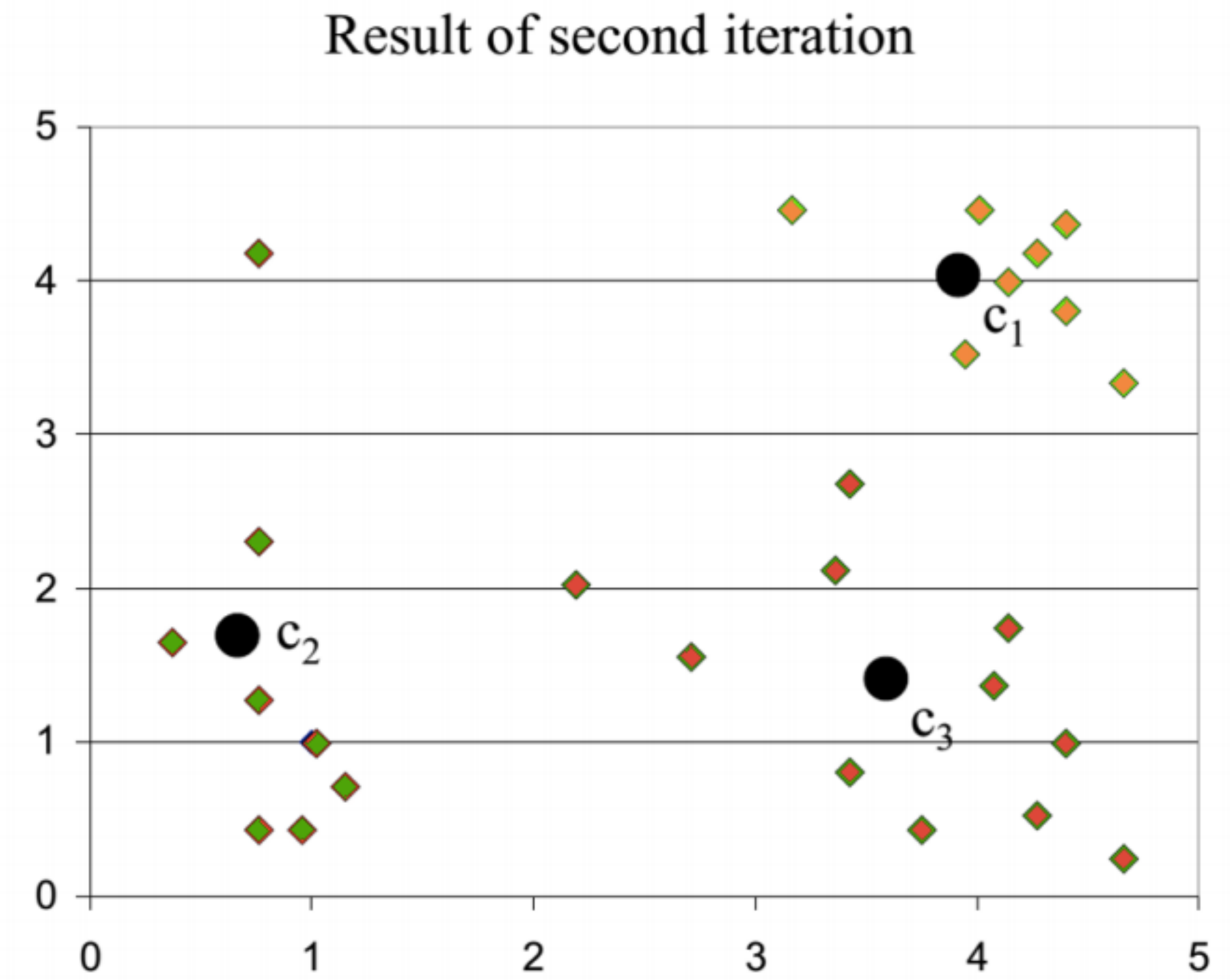


K-means Clustering : Example



Step2

K-means Clustering : Example



Step3

K-means Clustering : Strength

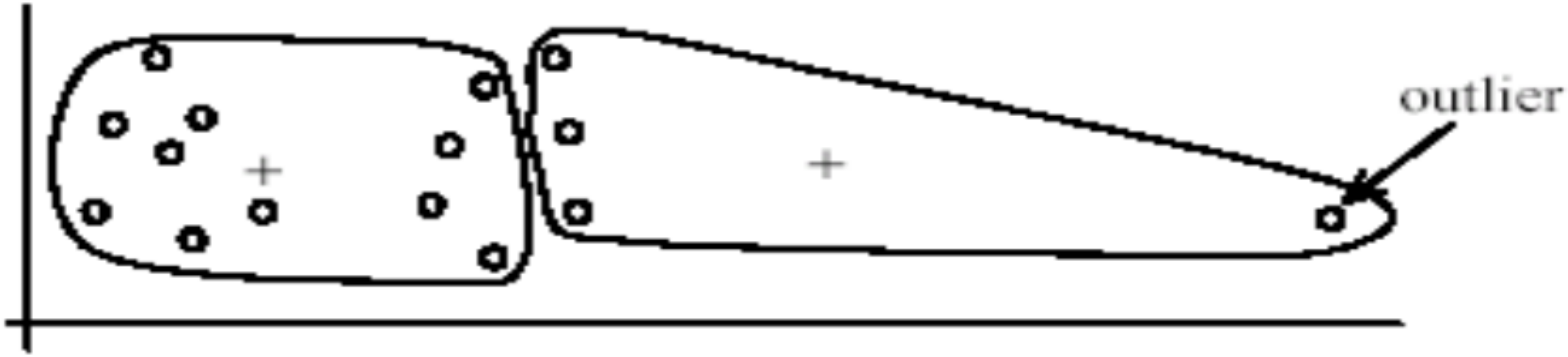
- Strengths:
  - Simple: easy to understand and to implement
  - Efficient: Time complexity: ***O(tkn)***,
    - where n is the number of data points,
    - k is the number of clusters, and
    - t is the number of iterations.
    - Since both k and t are small, k-means is considered an algorithm with linear time complexity.
- K-means is the most popular clustering algorithm.
- Note that: it terminates at a local optimum if SSE is used.
- The global optimum is hard to find due to complexity.

Weakness of K-means Clustering

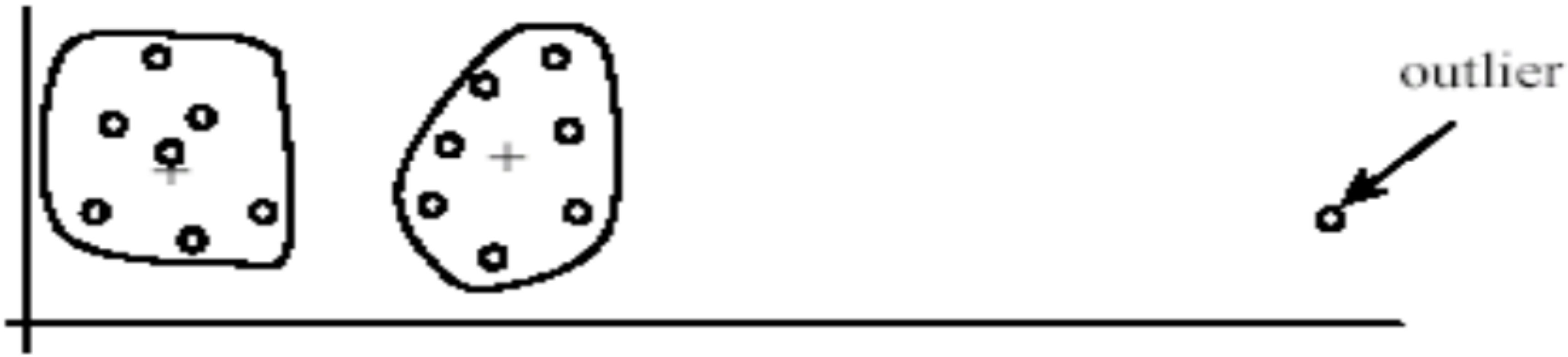
- The algorithm is only applicable if the mean is **defined**.
  - For categorical data, **k-mode** - the centroid is represented by most frequent values.
- The user needs to **specify** k.
- The algorithm is sensitive to **outliers**
  - Outliers are data points that are very far away from other data points.
  - Outliers could be errors in the data recording or some special data points with very different values.
- Output is sensitive to **initial means**

Weakness of K-means Clustering

Sensitive to outliers



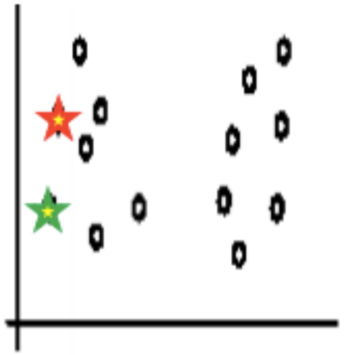
(A): Undesirable clusters



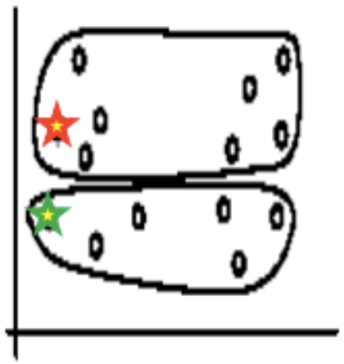
(B): Ideal clusters

Weakness of K-means Clustering

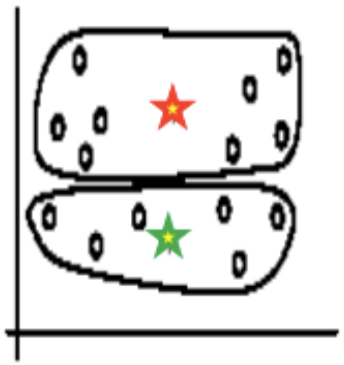
Sensitive to initial points



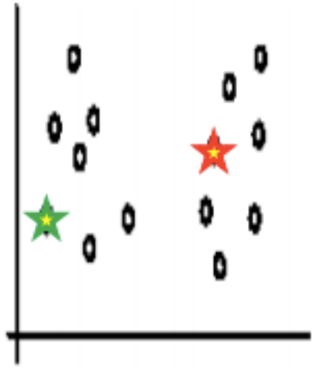
Random selection of seeds (centroids)



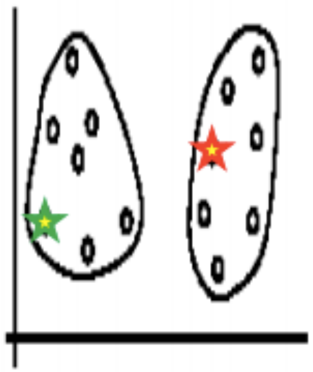
Iteration 1



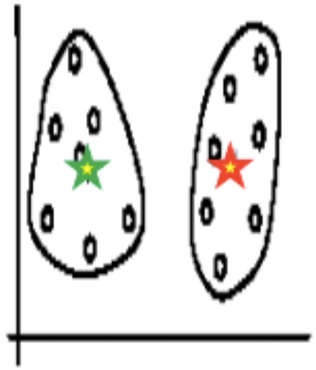
Iteration 2



Random selection of seeds (centroids)



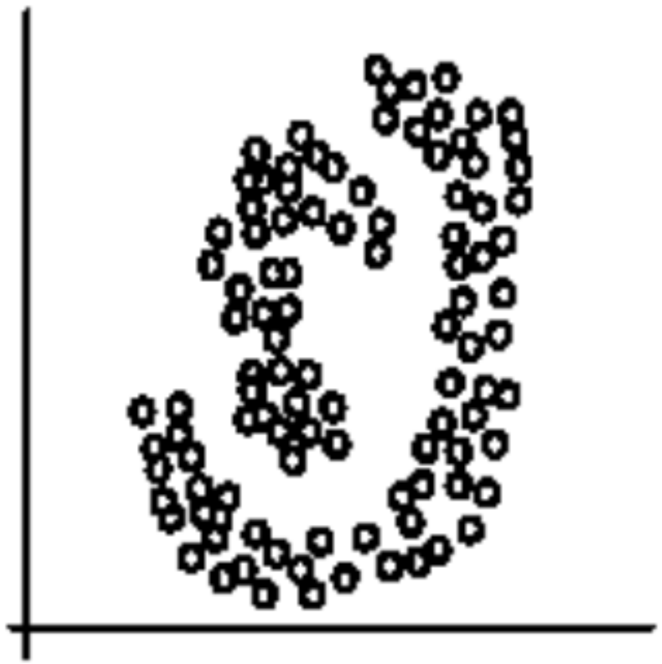
Iteration 1



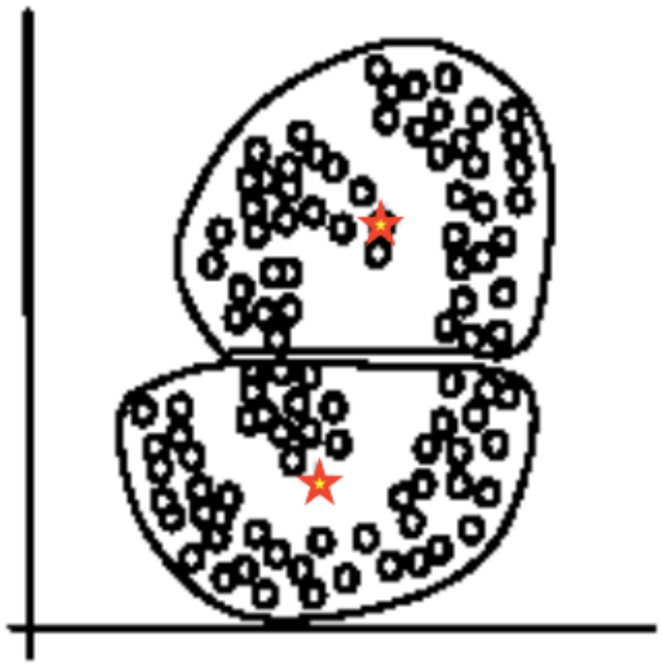
Iteration 2

Weakness of K-means Clustering

Not able to find non-hyper-ellipsoidal cluster



(A): Two natural clusters



(B): *k*-means clusters

Summery K-means Clustering

- Despite weaknesses, k-means is still the most popular algorithm due to its simplicity and efficiency
- No clear evidence that any other clustering algorithm performs better in general
- Comparing different clustering algorithms is a difficult task. No one knows the correct clusters!