INFDTA01-2 Data Science – Data mining

Giulia Costantini [office H.4.204]

What is this course about?

Data mining

- Explosive growth of data
  - terabytes of available data
  - automated data collection tools, database systems, web, …
  - major sources of abundant data: business, science, society, ...
- Information “hidden” in the data
  - human analysts take weeks to discover useful information

⇒ pressing need for the *automated analysis* of such massive data!

Lesson structure

- Brief presentation on the theory
- Time to study by yourself & work in pairs on the homework assignments
  - Keep a social behaviour
  - Exploit the teacher’s presence to ask questions and answer your doubts

Learning materials

- ***Data Smart***, John W. Foreman, Wiley, 2013
  - Chapters 2, 4, [6], 8, 9
  - Datasets for the practical assignments

- Slides of the lessons

Grading

More details on the modulewijzer

- *Written exam*

- *Practical examination: assignments + oral check*
  - Part 1: clustering
  - Part 2: genetic algorithms
  - Part 3: forecasting
  - Oral check: programming of some parts of data mining algorithms related to the three parts of the assignments

- *Final grade*
  - 100% practical examination...
  - … but the written exam <u>**MUST**</u> be sufficient to pass the course!!!

Practical examination

- Programming the assignments at home
  - in pairs or alone
  - up to 3 points
- Oral check
  - strictly individual
  - up to 7 points

- Admitted languages: Java, C#, Scala, F#
- Important date
  - <u>Lesson of week 8</u>
    - Check of your practical assignments & Oral check

More details on the modulewijzer

Program

- 8 weekly lessons
- Topics per week (draft)

Week	Topic
1	Introduction; Intro clustering
2	Clustering; Intro genetic algorithms
3	Genetic algorithms
4	Practicum; (optional) GA + regression case study
5	Forecasting (SES, DES)
6	Forecasting (TES)
7	Linear programming; outliers; summary/practicum
8	Check assignments + oral check

Today’s topics

- Introduction to data mining
- Introduction to clustering

Data mining intro

Data mining – Definitions

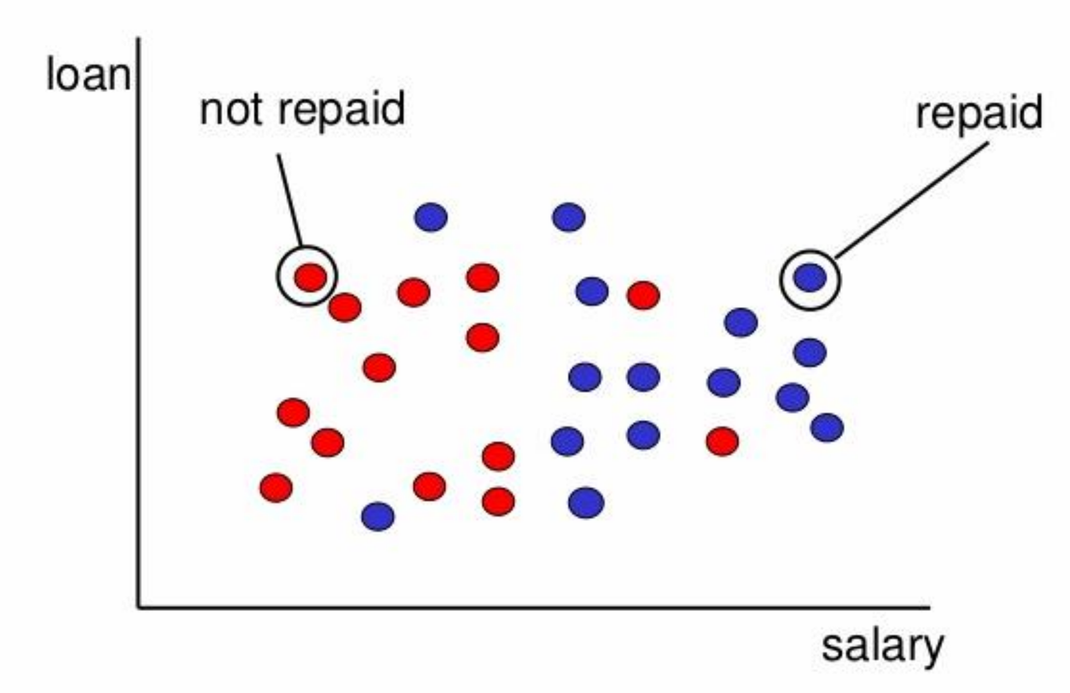
- Science of discovering structure and making predictions in (large) data sets
- Non-trivial extraction of implicit, previously unknown and useful information from data
- Exploration & analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns
- Process of semi-automatically analysing large databases to find patterns that are
    - Valid , Novel , Useful , Understandable

Data mining – Examples

- *In vitro fertilization*
  - Given: embryos described by 60 features
  - Problem: selection of embryos that will survive
  - Data: historical records of embryos and outcome
- *Credit assessment*
  - Given: a loan application
  - Problem: predict whether the bank should approve the loan
  - Data: records from other loans

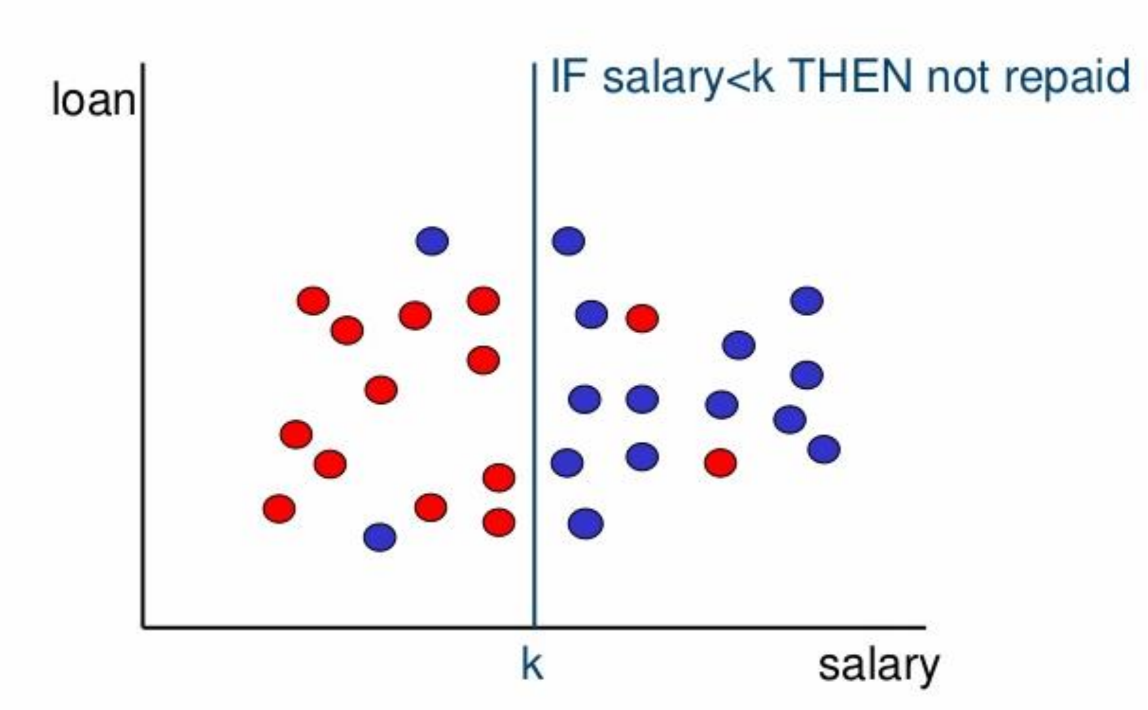
Data mining – Example

- Credit assessment



Data mining – Example

- Credit assessment

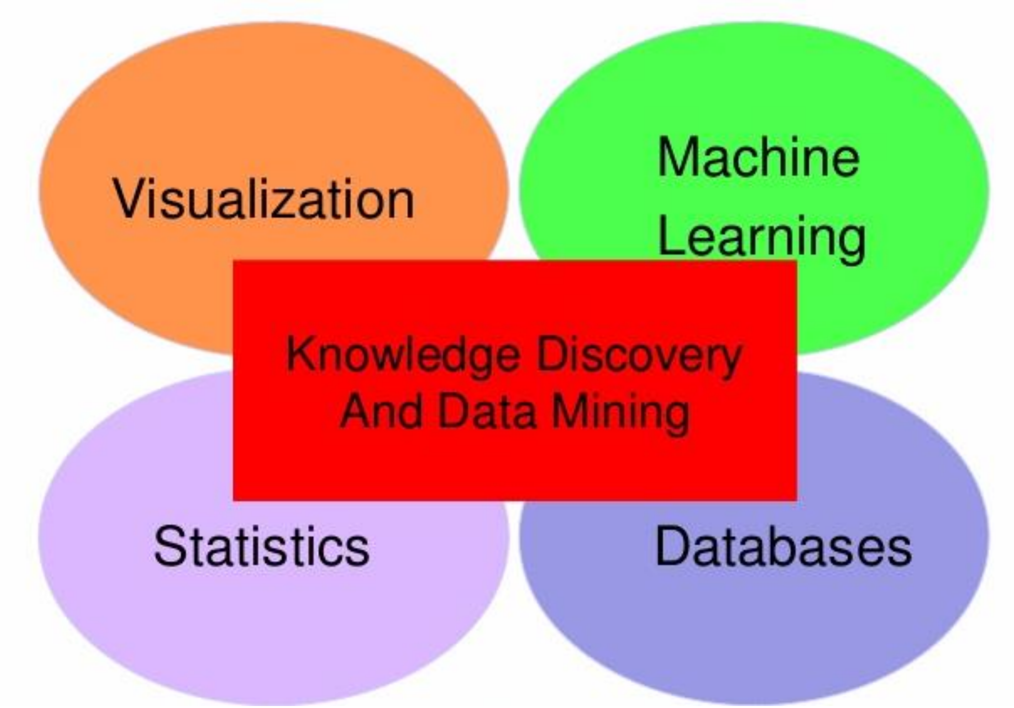


Data mining – Example

- Credit assessment
  - Valid?
    - the pattern has to be valid with respect to a certainty level (rule true for the 86%)
  - Novel?
    - the value k should be previously unknown and not obvious
  - Useful?
    - the pattern should provide information useful to the bank for assessing credit risk
  - Understandable?

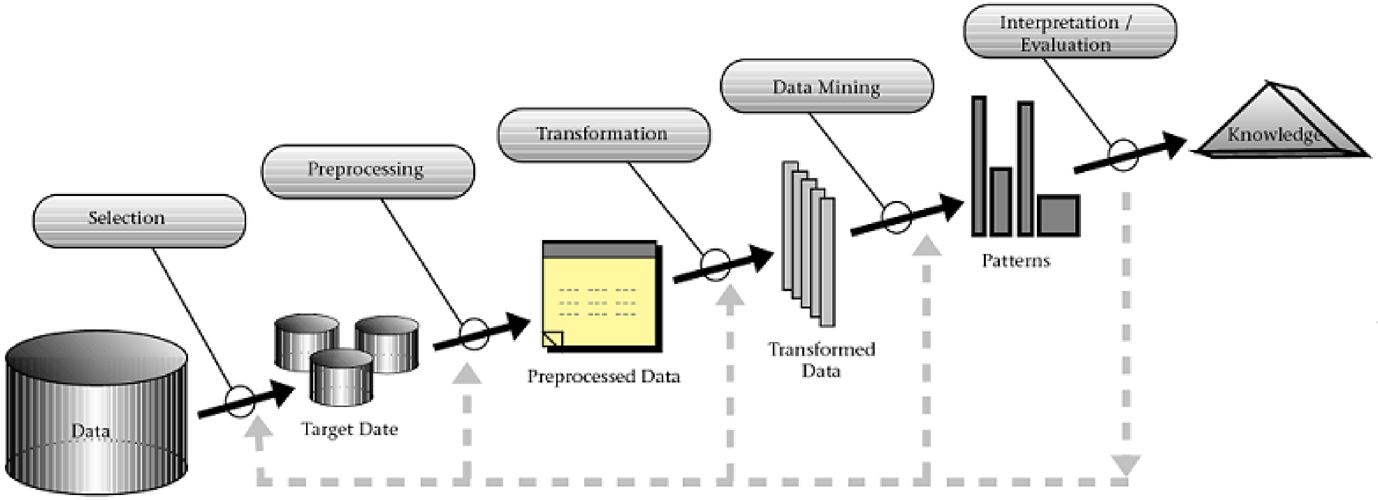
Data mining

- Related fields

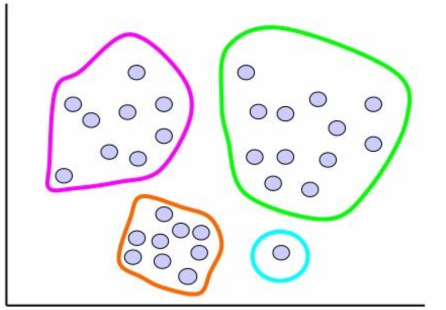


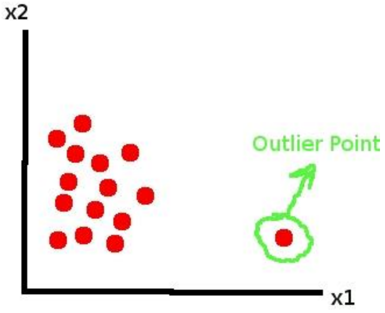
Data mining (DM) vs KDD

- *Data mining* ⇒ algorithms to extract patterns from data
- *KDD* ⇒ Knowledge Discovery from Data
  - overall process of discovering useful knowledge from data
  - DM focuses only on the application of some particular algorithms without the additional steps of the KDD process (like data cleaning, data reduction, visualization, etc…)

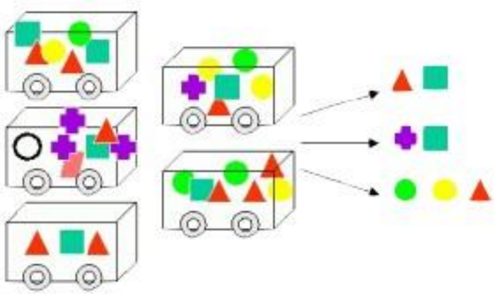


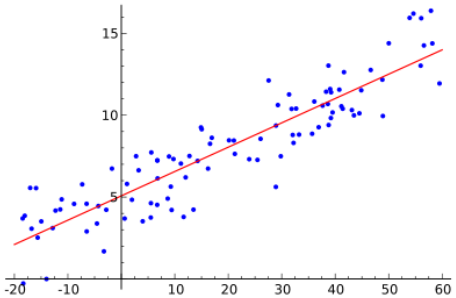
Data mining





- Major data mining tasks
  - Classification (predicting an item class)
  - **Clustering** (finding clusters in data)
  - Association rule (associations and/or correlation relationships)
  - **Estimation** (predicting a continuous value)
  - **Outlier analysis** (detect significant deviation from normal behaviour)
  - Trend and evolution analysis (**regression** analysis, sequential pattern mining, periodicity analysis, similarity-based analysis)





Data mining

- Classification
  - Finding models (functions) that describe and distinguish classes or concepts

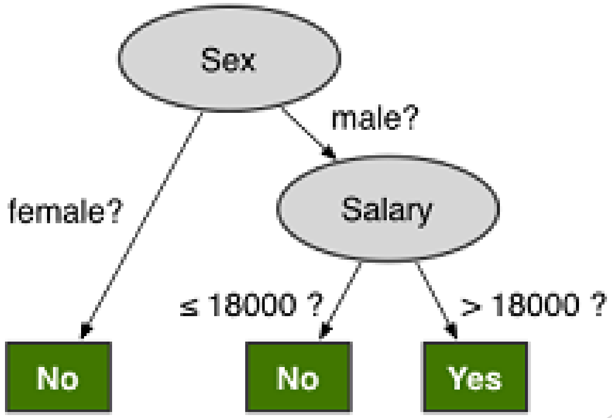
Classification Target

Attributes

Instances

Name	Salary	Sex	Age	Buy widget
Bloggs	15000	male	19	No
Jones	25000	male	33	Yes
Smit	23000	female	50	No
Smit	16000	male	40	No
Smit	200	male	10	No
Patel	30000	female	30	No
Steel	25000	male	23	Yes
Higgs	18000	female	55	No
Puggs	50000	male	57	Yes
Puggs	51000	female	57	No

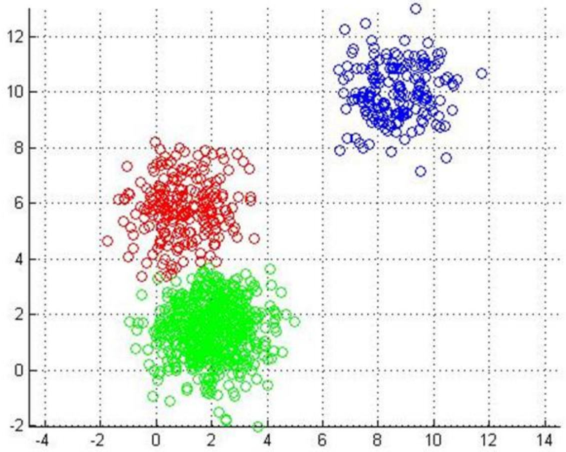
...

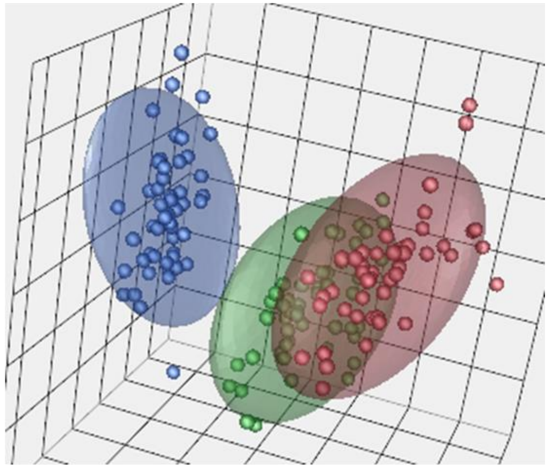


Clustering

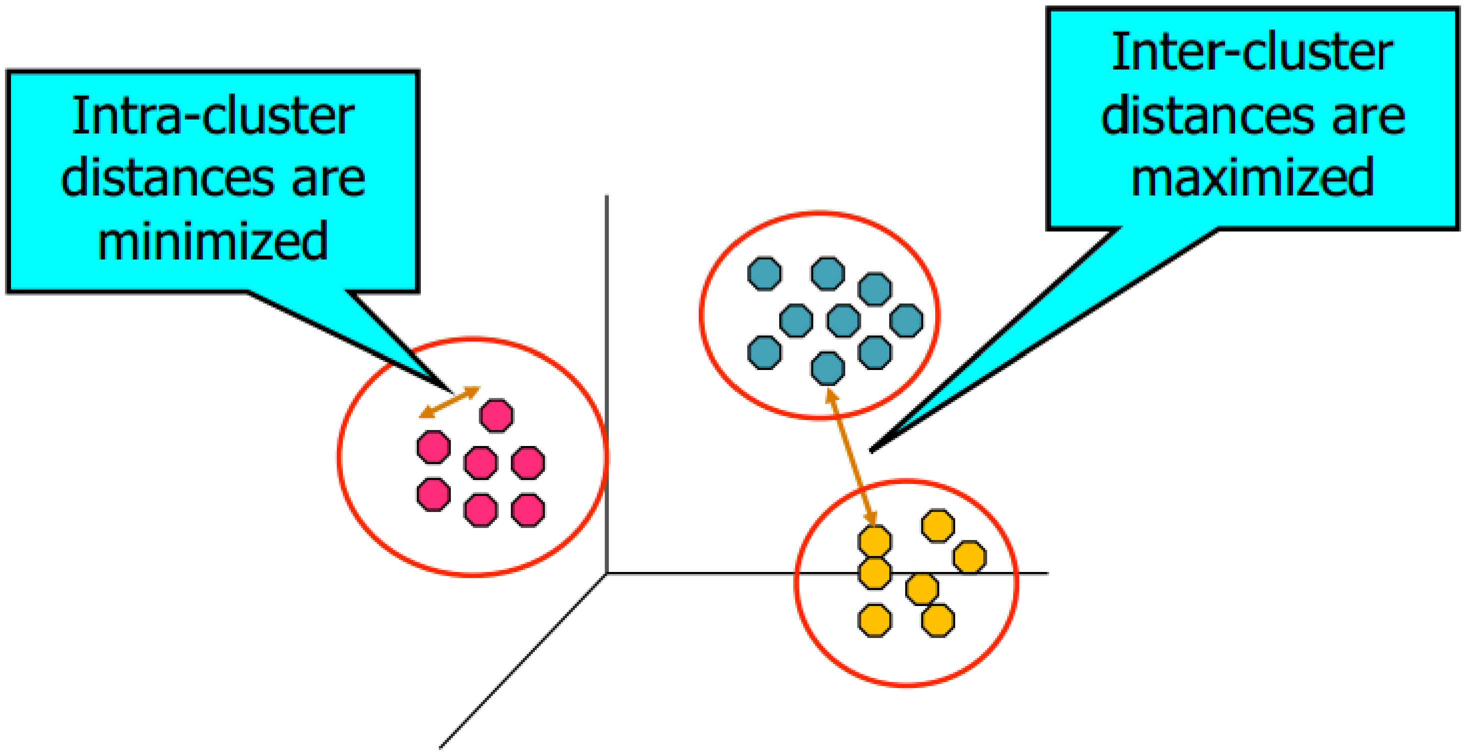
Clustering

- Finding a *structure* in a collection of unlabeled data
- Grouping a set of objects in such a way that…
  - objects in the same group (called a **cluster**) are more “similar” to each other than to those in other groups (clusters)





Clustering


Intra-cluster distances are minimized
Inter-cluster distances are maximized

Applications

- *Market segmentation*
  - discover distinct groups of customers and use this knowledge to develop targeted marketing programs
- *Biology*
  - classification of plants and animals given their features
- *City-planning*
  - identify groups of houses according to their house type, value and geographical location
- *Earthquake studies*
  - identify dangerous zones based on observed earthquake epicenters
- *WWW*
  - document classification; weblog clustering to discover groups of users with similar access patterns

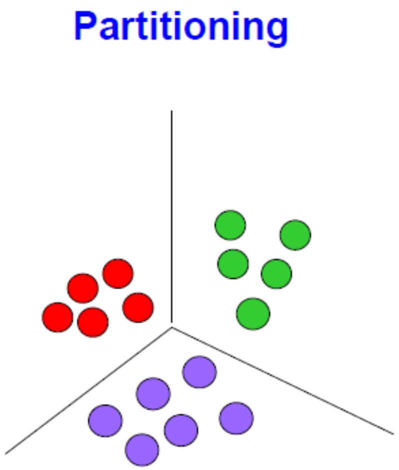
Clustering

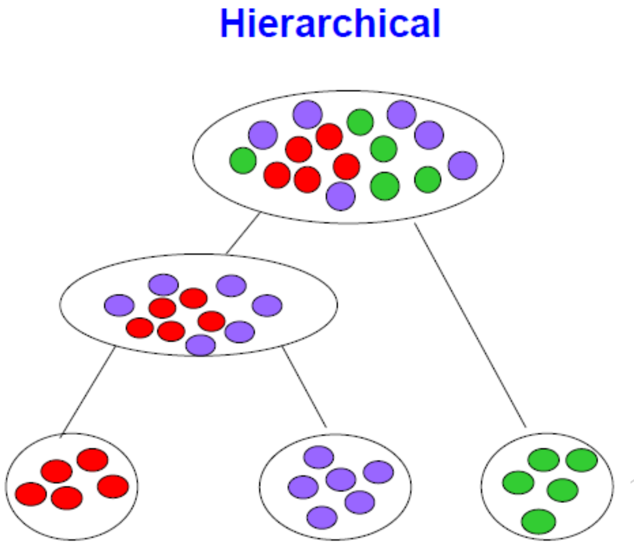
- General task to be solved
  - Can be achieved by many algorithms

- Partitioning approach (i.e., **k-means**)
- Hierarchical approach
- Density-based approach
- Grid-based approach
- Model-based
- Frequent-pattern based
- …

Clustering

- Partitional clustering
  - Objects are divided into non-overlapping subsets (clusters) such that each object is in exactly one subset
- Hierarchical clustering
  - A set of nested clusters organized as a hierarchical tree

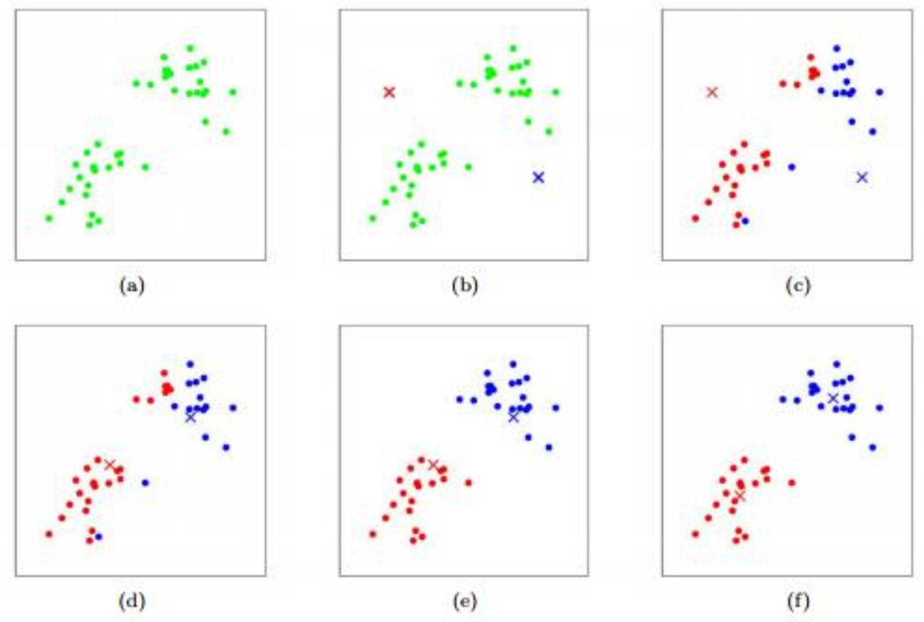




K-means clustering algorithm

- Goal: partition a set of observations into k clusters
  - each observation belongs to the cluster with the nearest mean/centroid, serving as a prototype of the cluster

- Iterative technique
  1) Initialization: choose a set of k initial means/centroids
  2) Assign each observation to the closest mean (i.e., cluster center)
  3) Recompute the cluster centers (as the mean of all the observations belonging to the cluster)
  4) Go back to step 2 and repeat



K-means clustering algorithm

- Steps

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

K-means clustering algorithm

- Pseudo-code

Randomly initialize K cluster centroids $\mu_1, \mu_2, \ldots, \mu_K \in \mathbb{R}^n$

Repeat {

    for i = 1 to m

        $c^{(i)}$:= index (from 1 to K) of cluster centroid closest to $x^{(i)}$

    for k = 1 to K

        μ_k := average (mean) of points assigned to cluster k

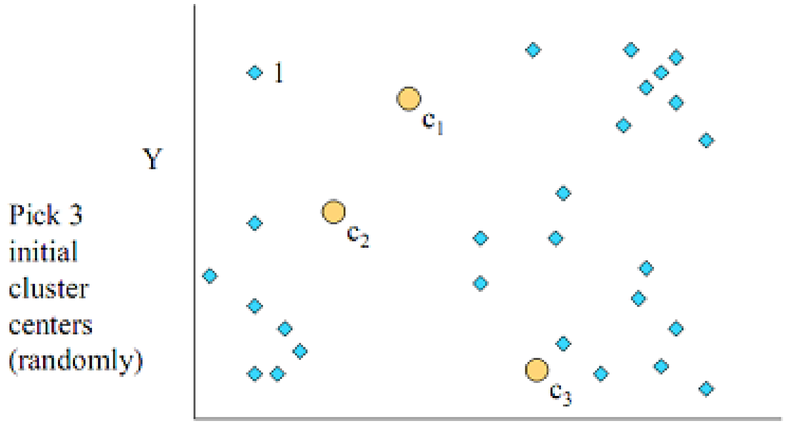
}

Assign the m observations to clusters

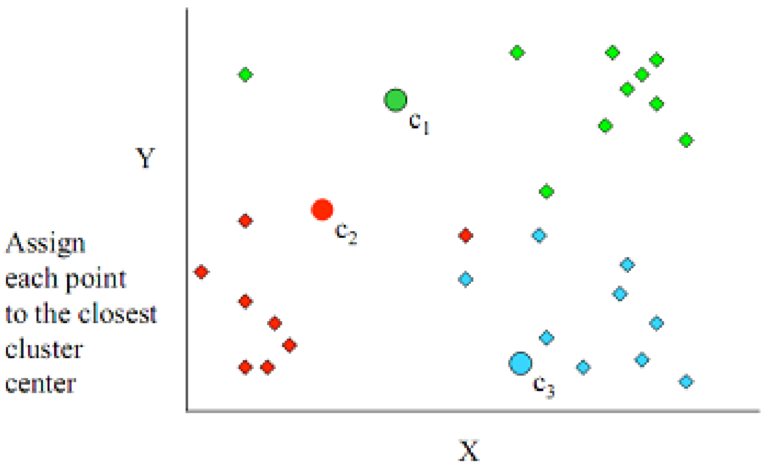
Update central points (clusters centers)

K-means clustering algorithm

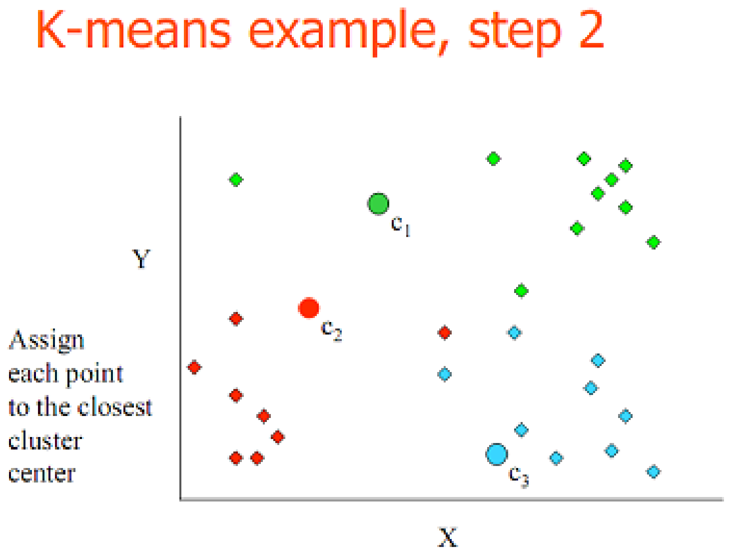
K-means example, step 1

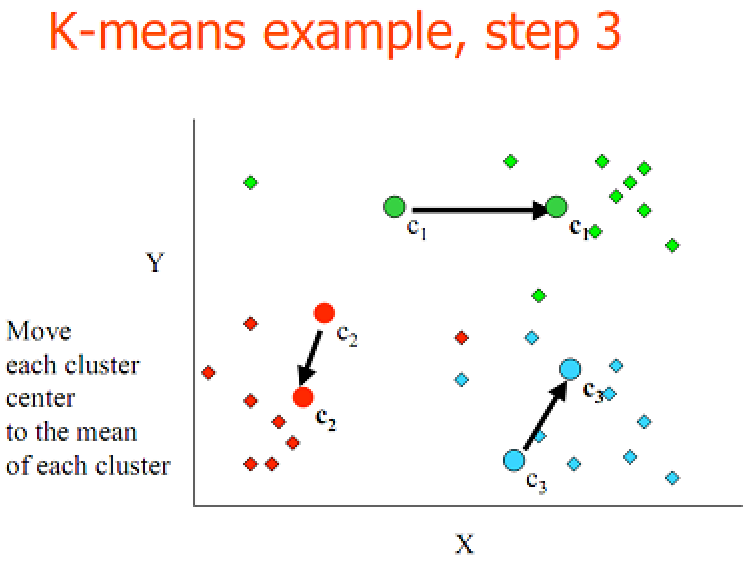


K-means example, step 2



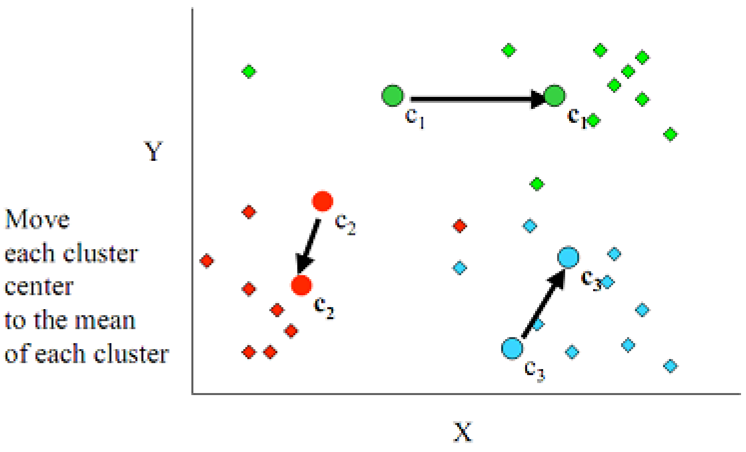
K-means clustering algorithm



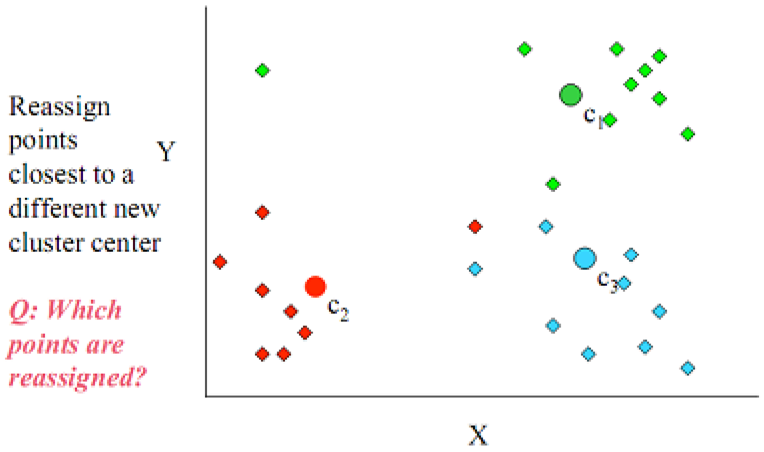


K-means clustering algorithm

K-means example, step 3

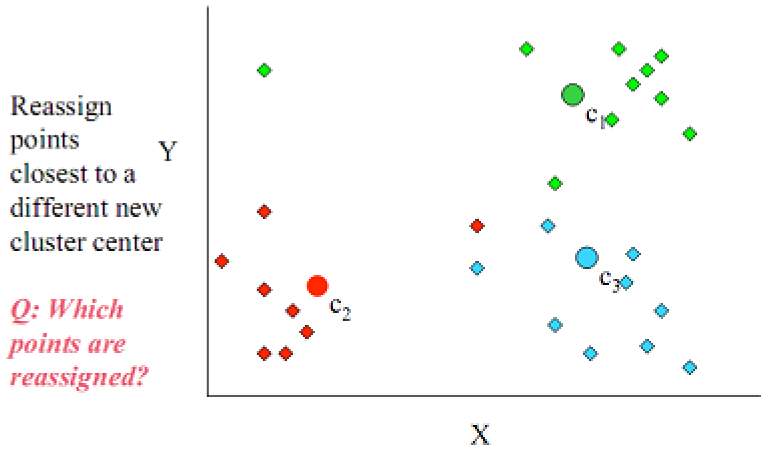


K-means example, step 4a

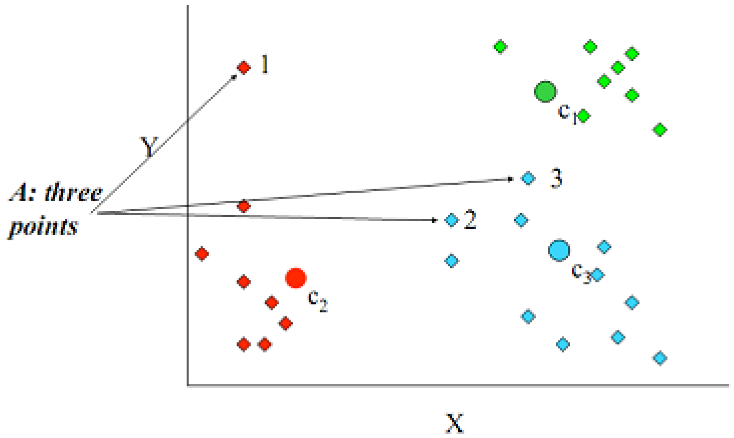


K-means clustering algorithm

K-means example, step 4a

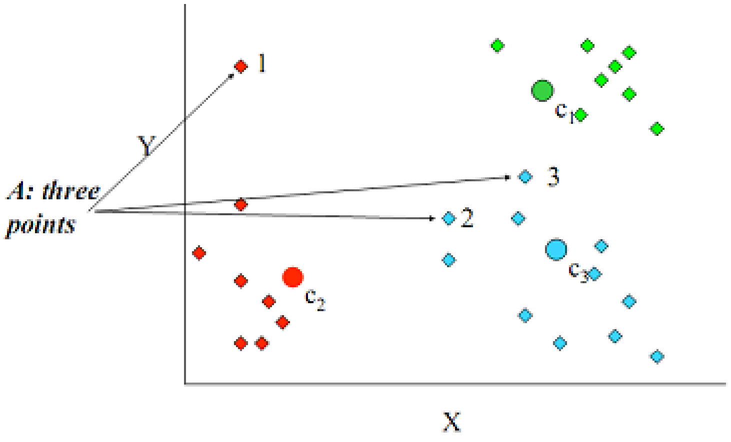


K-means example, step 4c

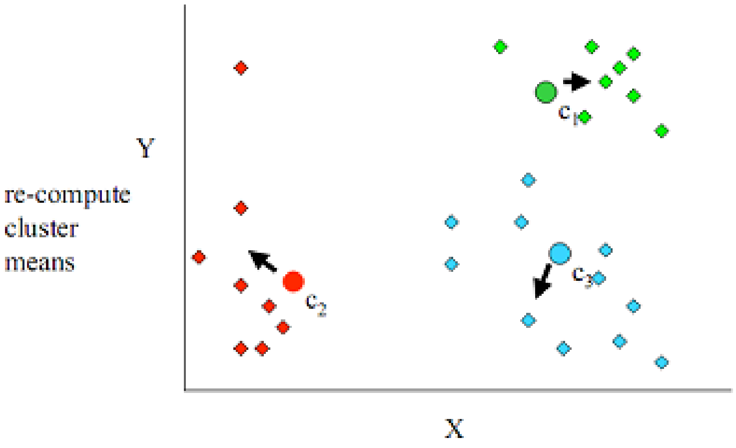


K-means clustering algorithm

K-means example, step 4c

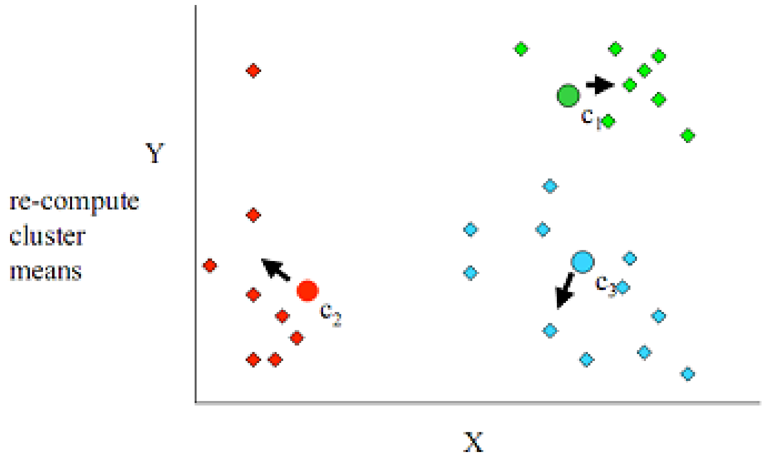


K-means example, step 4d

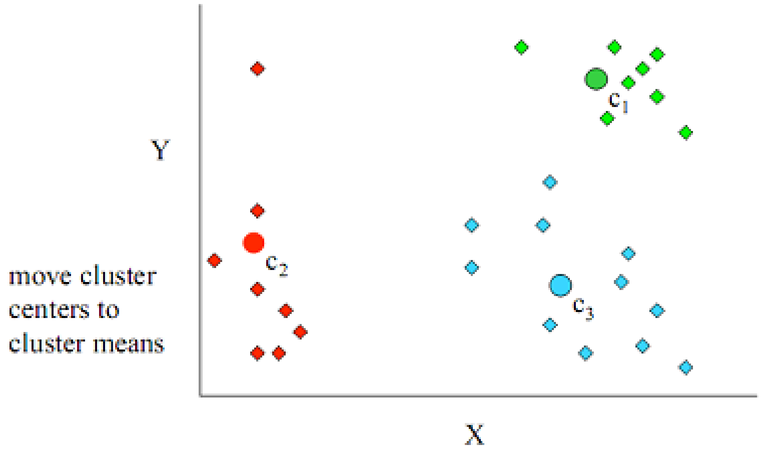


K-means clustering algorithm

K-means example, step 4d



K-means example, step 5

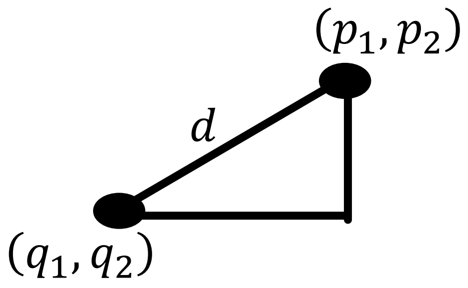


K-means issues

1. How to compute the distance between observations?
   - Many distances are possible
   - Euclidean distance

$$2\text{–dim}: \quad \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}$$

$$n\text{–dim}: \quad \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$



K-means issues

2. What is a good value for k?
   - Input parameter of the algorithm
   - An inappropriate choice may yield poor results
   - More on this next week!

K-means issues

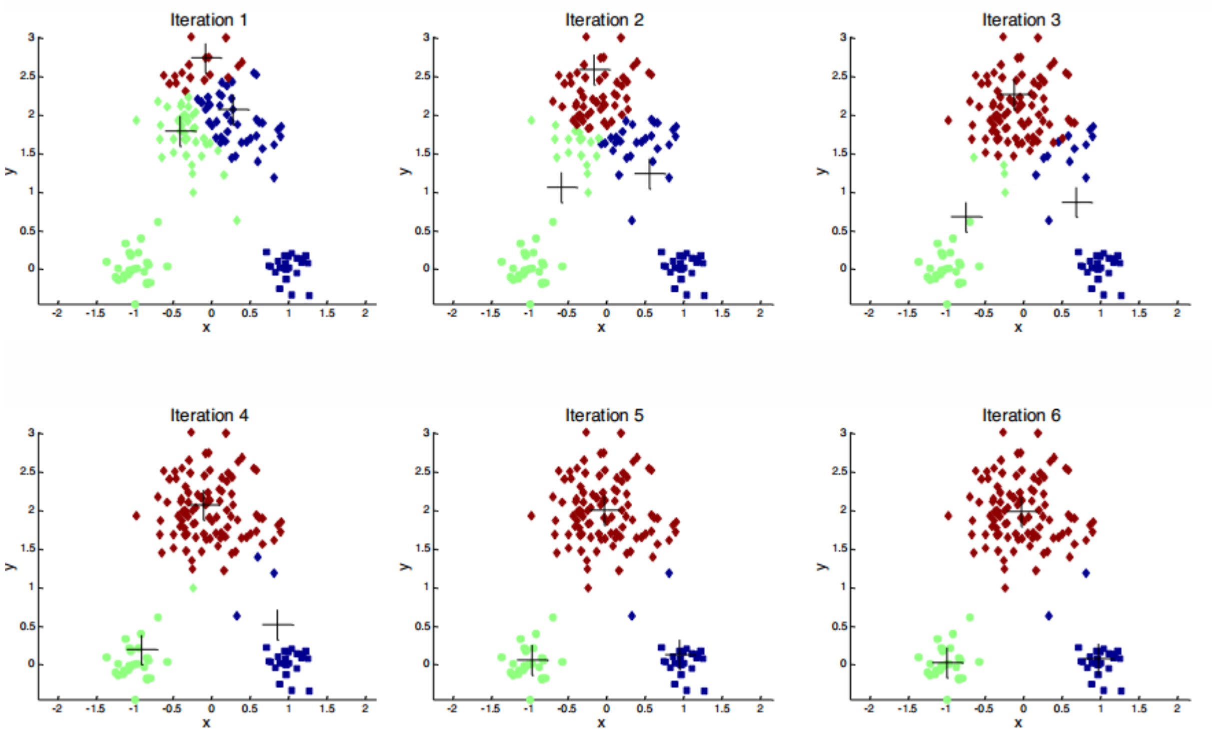
3. When to terminate the algorithm?
   - When a maximum number of iterations is reached
   - When the centroids stop changing

K-means issues

4. How to choose the first set of k means?
   - Random points inside the space
   - Forgy method
     - Choose randomly k observations from the dataset
   - Random Partition method
     - Divide randomly the observations in clusters
     - The first k means are the centroids of the randomly made clusters

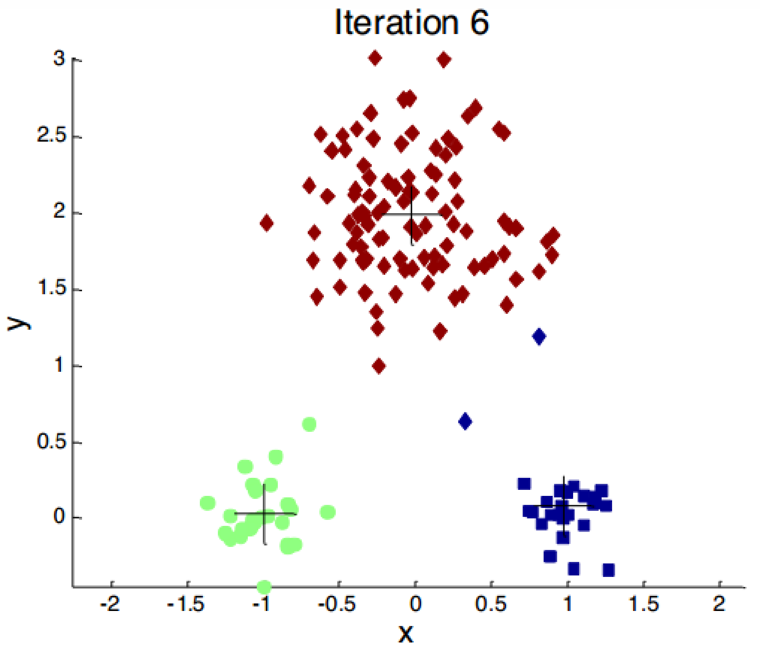
K-means issues

- Good choice of initial cluster centers



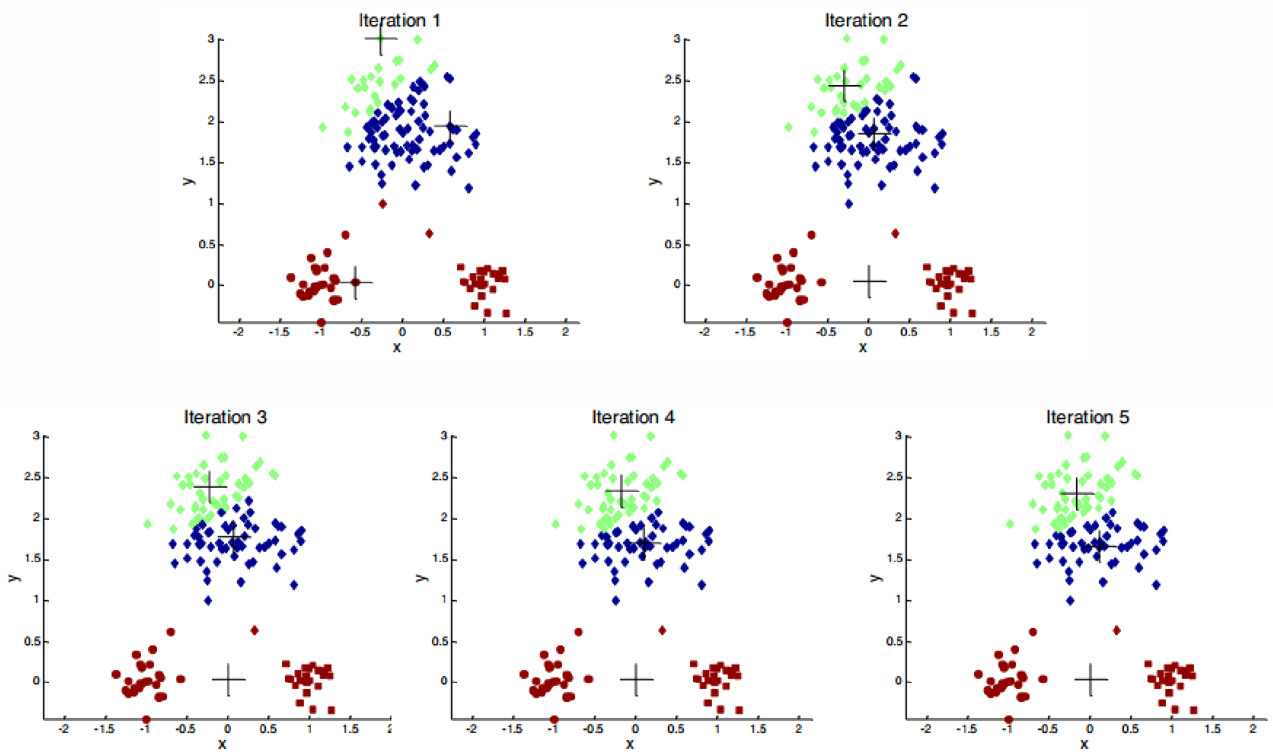
K-means issues

- Good choice of initial cluster centers



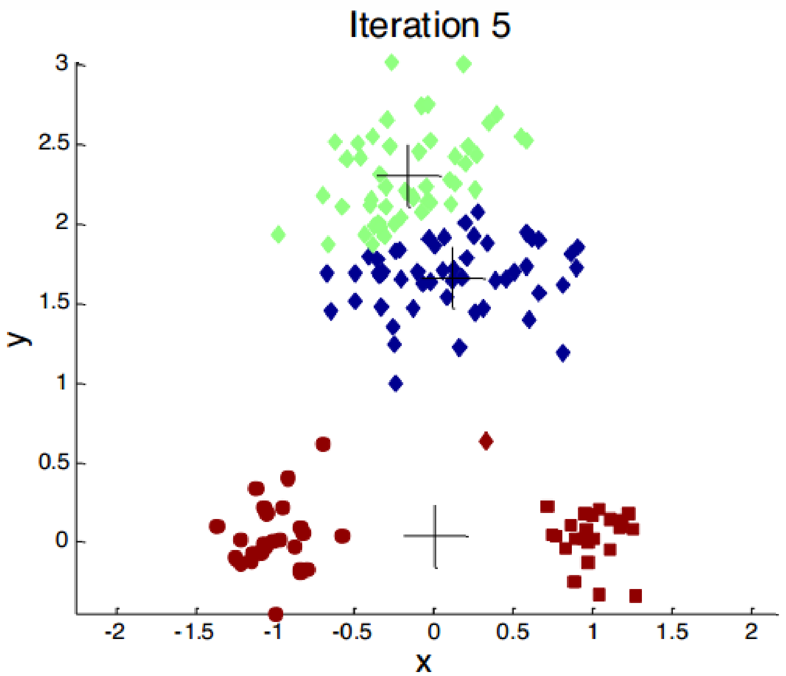
K-means issues

- Bad choice of initial cluster centers



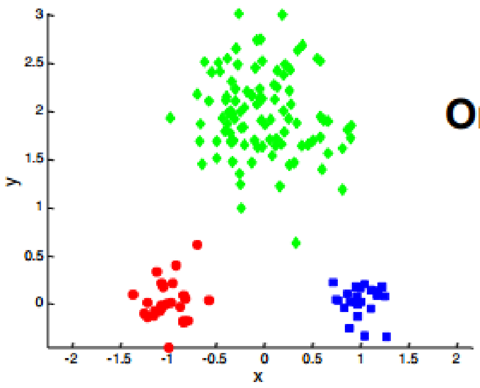
K-means issues

- Bad choice of initial cluster centers

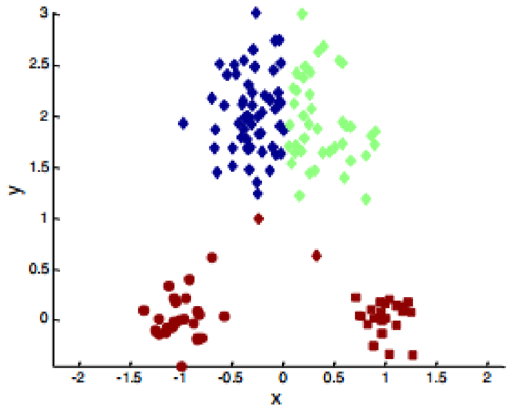


K-means issues

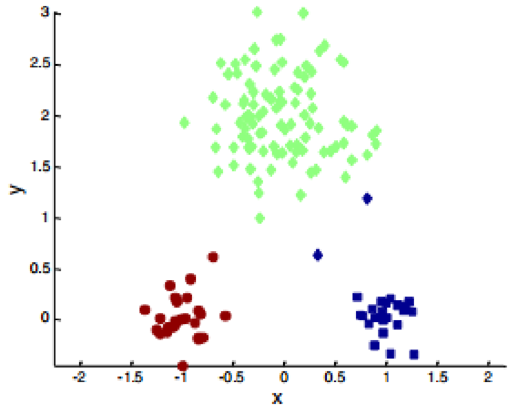
- Initial cluster centers: very important choice!
  - the quality of the result depends greatly on it



Original Points



Sub-optimal Clustering



Optimal Clustering

K-means issues

- Initial cluster centers: how to find the best?
  - *Multiple runs* (helps a bit)
  - Use another clustering method (hierarchical?) to determine initial centroids
  - Select more than k initial centroids and then select among these (the most widely separated)
  - Post-processing
  - Bisecting k-means

K-means issues

Multiple runs

- Given k, to determine the “best” clustering solution
  - Repeat the k-means algorithm many times (i.e., 50-100)
    - Each time, random choice of initial cluster centers
  - For each run, compute a measure of the *global error*
  - Choose the solution with *lowest error*

- Most common measure of error
  - **SSE** → how closely related are objects in a cluster

K-means issues

- SSE = Sum of Squared Errors
  - Sum of the squares of the error of each point
  - Error of a point = *distance* to the nearest cluster center

$$SSE = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(c_i, x)$$

where

- x = data point in the cluster C_i
- c_i = center of cluster C_i

Homework

- Study **slides** week 1
- Read **Chapter 2** of the book (*until page 53*)
- Download **datasets** for part 1 of the assignment
  - The complete one: “WineKMC.xlsx” from http://eu.wiley.com/WileyCDA/WileyTitle/productCd-111866146X.html → Downloads section → Chapter 2
  - The already preprocessed in csv format (“Winedata.csv”) from N@tschool
- Start the implementation of **part 1**… ?
  - Read description of the assignment (available on N@tschool)
  - We will also talk about it together during the next lesson