Machine Learning: Theory and Applications

Arnak S. Dalalyan
American University of Armenia, Yerevan
April 6, 2016

ENSAE ParisTech

Short CV and contacts

arnak-dalalyan.fr

ENSAE ParisTech

Short CV and contacts

arnak-dalalyan.fr

1979 Born in Yerevan (tramvi park)

ENSAE ParisTech

Short CV and contacts

arnak-dalalyan.fr

1979 Born in Yerevan (tramvi park)

1997 Bachelor degree in Mathematics, YSU

1999 Master degree in Probability and Statistics, Paris 6 University

ENSAE ParisTech

Short CV and contacts

arnak-dalalyan.fr

1979 Born in Yerevan (tramvi park)

1997 Bachelor degree in Mathematics, YSU

1999 Master degree in Probability and Statistics, Paris 6 University

2001 PhD in Statistics, University of Le Mans (France)

ENSAE ParisTech

Short CV and contacts

arnak-dalalyan.fr

1979 Born in Yerevan (tramvi park)

1997 Bachelor degree in Mathematics, YSU

1999 Master degree in Probability and Statistics, Paris 6 University

2001 PhD in Statistics, University of Le Mans (France)

2007 Habilitation in Statistics and Machine Learning, Paris 6 University

ENSAE ParisTech

Short CV and contacts

arnak-dalalyan.fr

1979 Born in Yerevan (tramvi park)

1997 Bachelor degree in Mathematics, YSU

1999 Master degree in Probability and Statistics, Paris 6 University

2001 PhD in Statistics, University of Le Mans (France)

2007 Habilitation in Statistics and Machine Learning, Paris 6 University

Now

- Professor of Statistics and Applied Mathematics, ENSAE ParisTech
- Head of the Master in Data Science
- Deputy chair of the Center for Data Science Paris-Saclay
- Programme committee of the COLT (conf. on Learning Theory) and NIPS (Neural Information and Processing Systems), Associate Editor of EJS, JSPI, SISP and JSS.

ENSAE ParisTech

What is Machine Learning ?

The emphasis of machine learning is on devising automatic methods that perform a given task based on what was experienced in the past.

Here are several examples :

- optical character recognition : categorize images of handwritten characters by the letters represented
- topic spotting : categorize news articles (say) as to whether they are about politics, sports, entertainment, etc.
- medical diagnosis : diagnose a patient as a sufferer or non-sufferer of some disease
- customer segmentation : dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing (such as gender, interests, spending habits)
- fraud detection : identify credit card transactions (for instance) which may be fraudulent in nature

ENSAE ParisTech

Typical tasks of machine learning

Supervised learning :

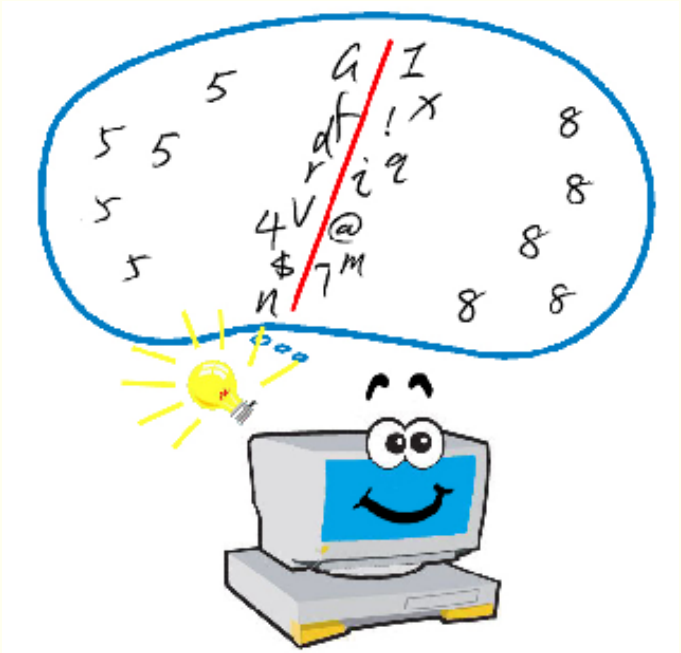
- Prediction : use historical data for predicting future values
- Classification : identifying to which of a set of classes a new observation belongs, on the basis of a training set of data containing observations whose class membership is known
- Ranking : construction of ranking models for information retrieval systems

Unsupervised learning :

- Outlier detection : detecting a few observations that deviate so much from other observations as to arouse suspicion that it was generated by a different mechanism
- Clustering : grouping a set of objects into homogeneous groups based on some similarity measure
- Dimensionality reduction : mapping of data to a lower dimensional space such that uninformative variance in the data is discarded

ENSAE ParisTech

Clustering Methods in Unsupervised Learning



Outline

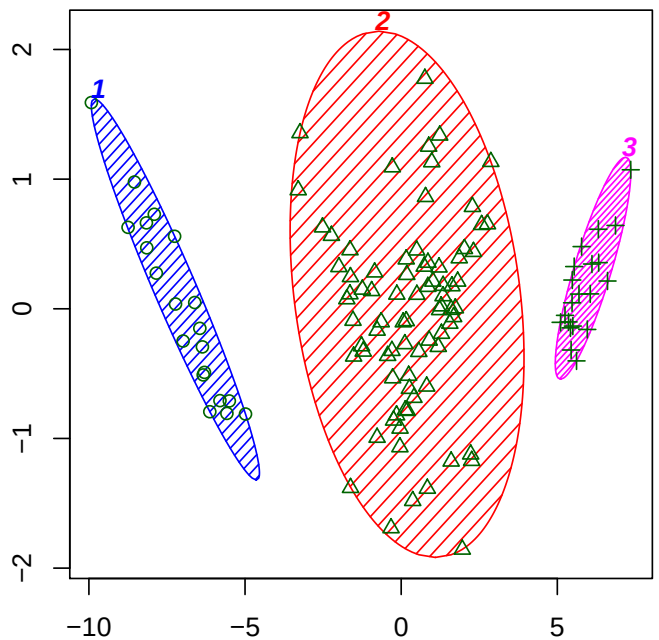
- Introduction to clustering
- K-means and partitioning around medoids
  - Theoretical background
  - R code
  - An example
- Gaussian mixtures and EM algorithm
  - Theoretical background
  - R code
  - An example
- Hierarchical Clustering
  - Theoretical background
  - R code
  - An example

ENSAE ParisTech

Introduction to clustering

Main objective

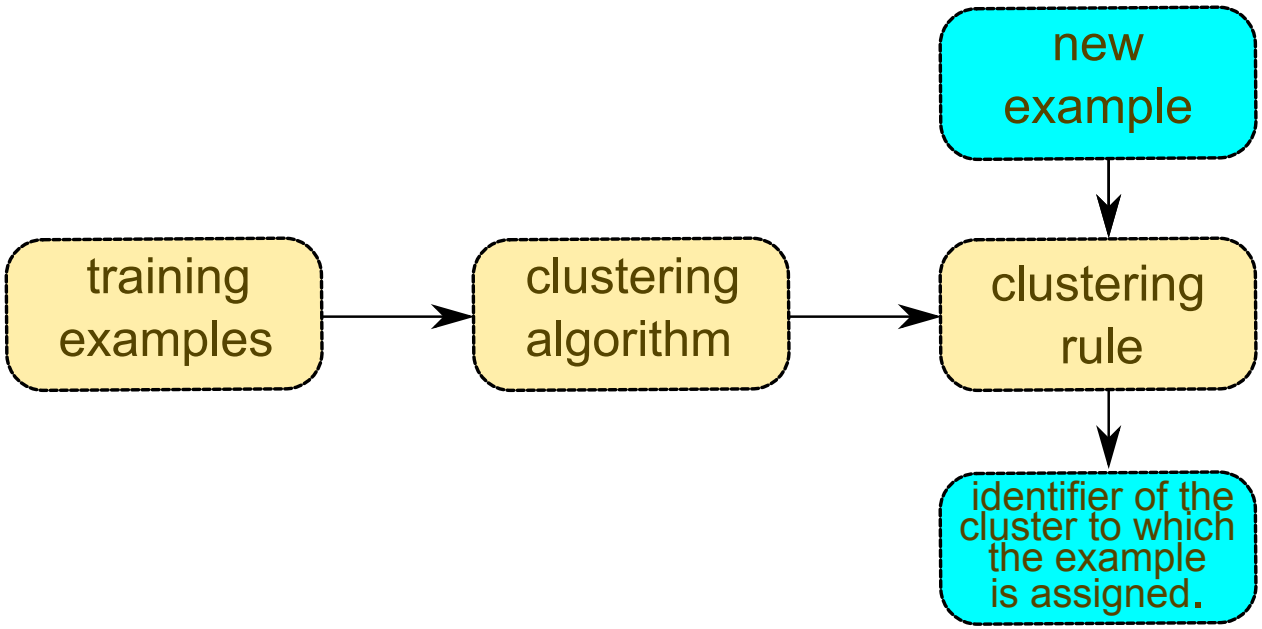
The goal of clustering analysis is to find high-quality clusters such that the inter-cluster similarity is low and the intra-cluster similarity is high.



ENSAE ParisTech

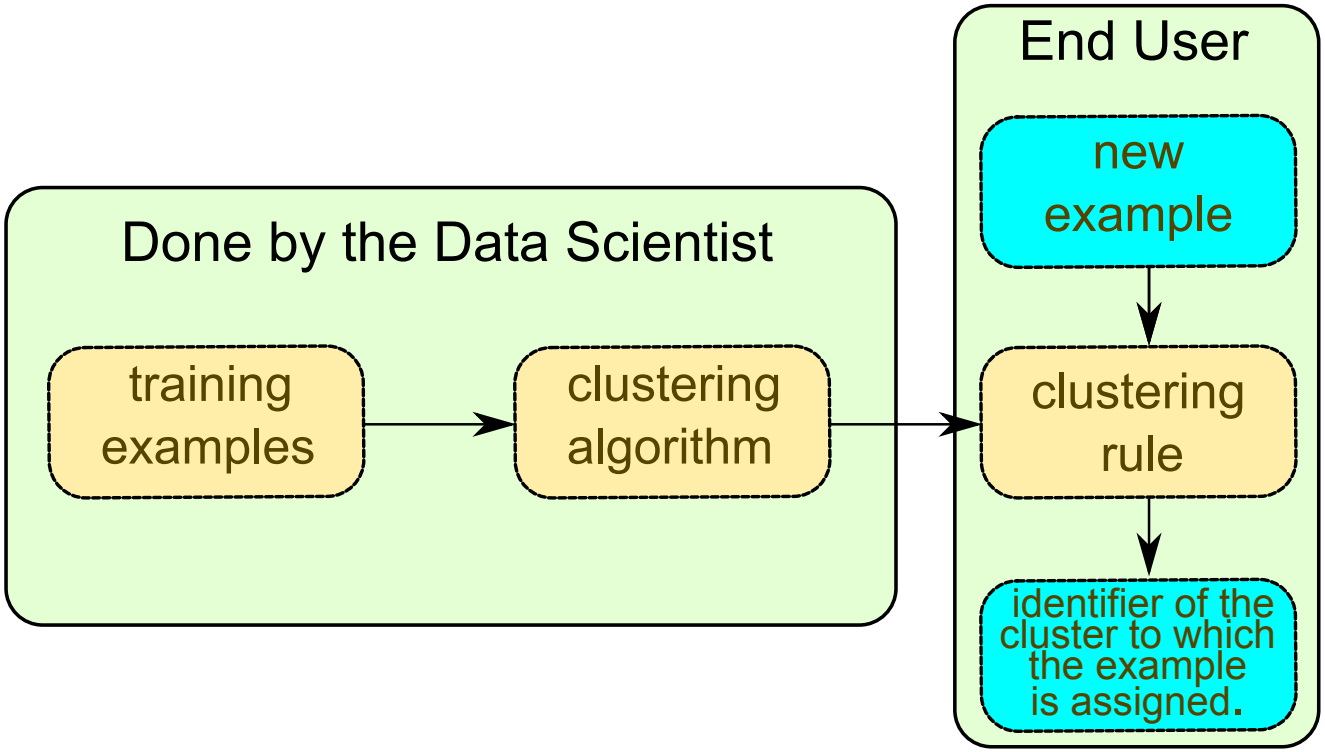
Introduction to clustering

The general diagram



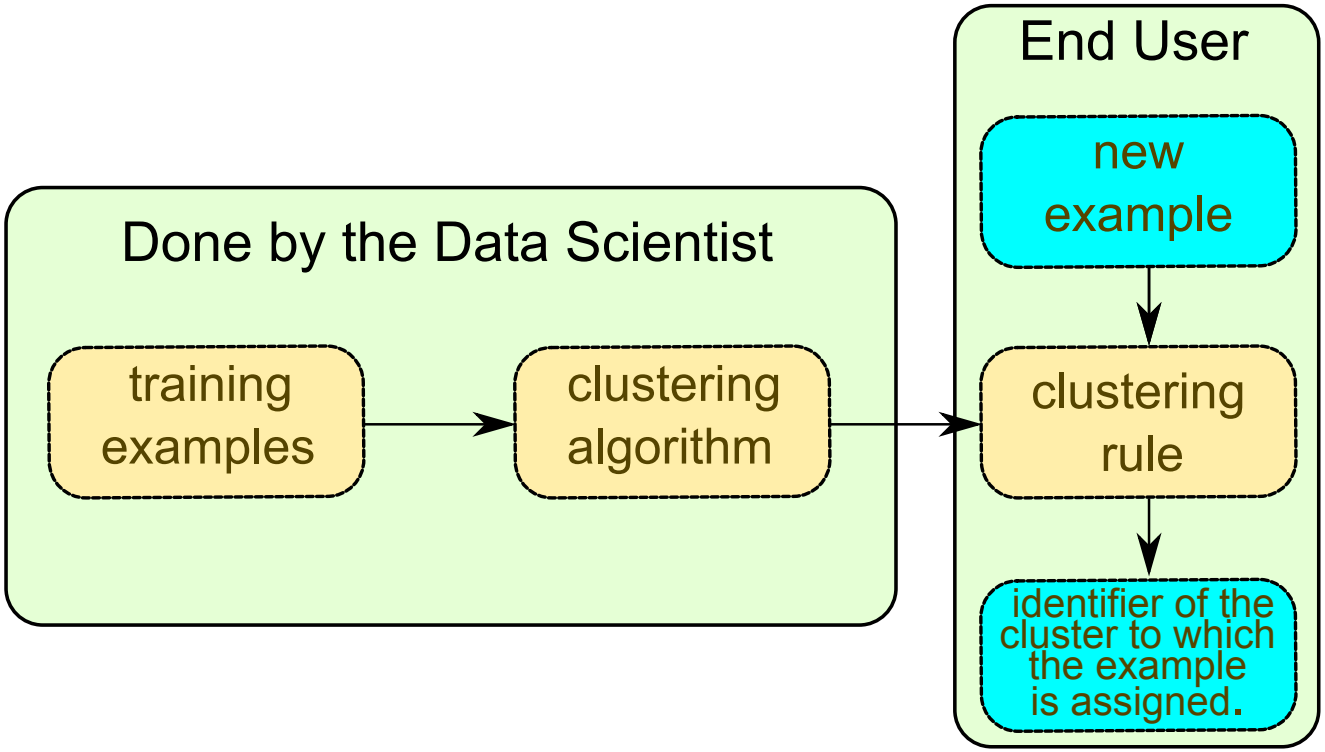
Introduction to clustering

The general diagram



Introduction to clustering

The general diagram



In general, there is no objective measure of quality for a clustering algorithm. The satisfaction of the end-user is perhaps the best manner of evaluation.

K-means clustering

Notation

- We are given n examples represented as a p-vector with real-values entries :

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^p, \qquad \mathbf{X}_i = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{bmatrix}.$$

- We denote by $\|\mathbf{X}_i - \mathbf{X}_j\|$ the (Euclidean) distance between two examples :

$$\|\mathbf{X}_i - \mathbf{X}_j\|^2 = \sum_{\ell=1}^{p} (x_{i,\ell} - x_{j,\ell})^2$$

- For a set G, subset of $\{1, \ldots, n\}$, we denote by $\overline{\mathbf{X}}_G$ the average of the examples $\mathbf{X}_i$ corresponding to $i \in G$, that is

$$\overline{\mathbf{X}}_G = \frac{1}{|G|} \sum_{i \in G} \mathbf{X}_i.$$

ENSAE ParisTech

K-means clustering

Definition of the method

- Main idea : A good clustering algorithm divides the sample into K groups such that the variance within each group is mall.
- Mathematically speaking, this corresponds to solving with respect to $G_1, \ldots, G_K \subset \{1, \ldots, n\}$ and $\mathbf{C}_1, \ldots, \mathbf{C}_K \in \mathbb{R}^p$ the following optimisation problem :

$$\min_{G_1,\ldots,G_K} \min_{\mathbf{C}_1,\ldots,\mathbf{C}_K} \underbrace{\sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2}_{\Psi(G_{1:K}, \mathbf{C}_{1:K})},$$

where $G_1, \ldots, G_K$ runs over all possible partitions of $\{1, \ldots, n\}$.

- Important remark : the minimization should be done for a fixed value of K (prescribed number of clusters), otherwise the solution is obvious (and meaningless), isn't it ?

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.
- Local minimum : finding a local minimum is easy.

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.
- Local minimum : finding a local minimum is easy.

  Step 1 For fixed centers $\mathbf{C}_{1:K}$ the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $G_{1:K}$ can be easily computed :

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.
- Local minimum : finding a local minimum is easy.

  Step 1 For fixed centers $\mathbf{C}_{1:K}$ the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $G_{1:K}$ can be easily computed :

  ▶ G_k contains all the points $\mathbf{X}_i$ for which the closest point in the set $\{\mathbf{C}_1, \ldots, \mathbf{C}_K\}$ is $\mathbf{C}_k$.

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.
- Local minimum : finding a local minimum is easy.

  Step 1 For fixed centers $\mathbf{C}_{1:K}$ the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $G_{1:K}$ can be easily computed :
  - G_k contains all the points $\mathbf{X}_i$ for which the closest point in the set $\{\mathbf{C}_1, \ldots, \mathbf{C}_K\}$ is $\mathbf{C}_k$.

  Step 2 For fixed groups $G_{1:K}$, the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $\mathbf{C}_{1:K}$ can be easily computed :

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.
- Local minimum : finding a local minimum is easy.

  Step 1 For fixed centers $\mathbf{C}_{1:K}$ the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $G_{1:K}$ can be easily computed :
  - G_k contains all the points $\mathbf{X}_i$ for which the closest point in the set $\{\mathbf{C}_1, \ldots, \mathbf{C}_K\}$ is $\mathbf{C}_k$.

  Step 2 For fixed groups $G_{1:K}$, the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $\mathbf{C}_{1:K}$ can be easily computed :
  - $\mathbf{C}_k = \overline{\mathbf{X}}_{G_k}$.

K-means clustering

How to minimize the cost ?

- Cost function : $\Psi(G_{1:K}, \mathbf{C}_{1:K}) = \sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|^2$ is hard to minimize. It is a combinatorial optimization problem which is known to be NP-hard.
- Local minimum : finding a local minimum is easy.

  Step 1 For fixed centers $\mathbf{C}_{1:K}$ the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $G_{1:K}$ can be easily computed :
  - G_k contains all the points $\mathbf{X}_i$ for which the closest point in the set $\{\mathbf{C}_1, \ldots, \mathbf{C}_K\}$ is $\mathbf{C}_k$.

  Step 2 For fixed groups $G_{1:K}$, the minimizer of $\Psi(G_{1:K}, \mathbf{C}_{1:K})$ w.r.t. $\mathbf{C}_{1:K}$ can be easily computed :
  - $\mathbf{C}_k = \overline{\mathbf{X}}_{G_k}$.

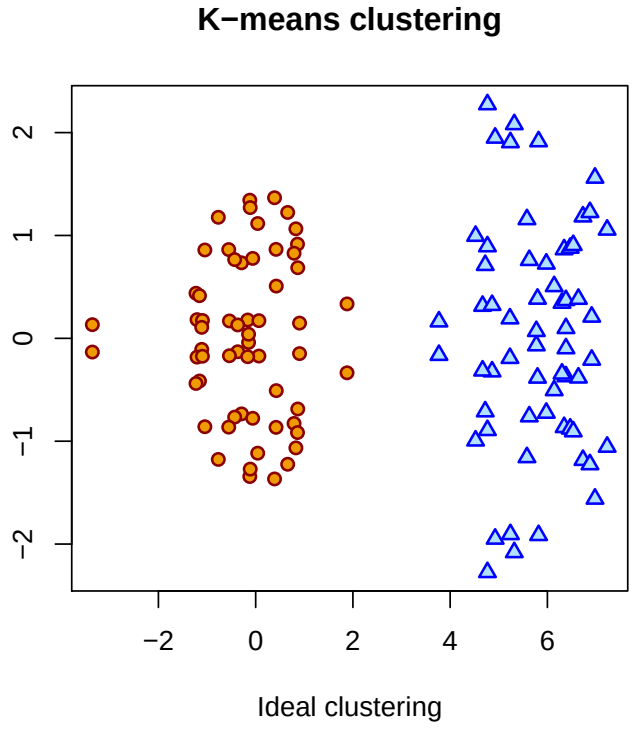
- Iterative algorithm : **initialize at some $\mathbf{C}_{1:K}$ and then alternate between the two aforementioned steps until the convergence.**

ENSAE ParisTech

K-means clustering

Alternating minimization

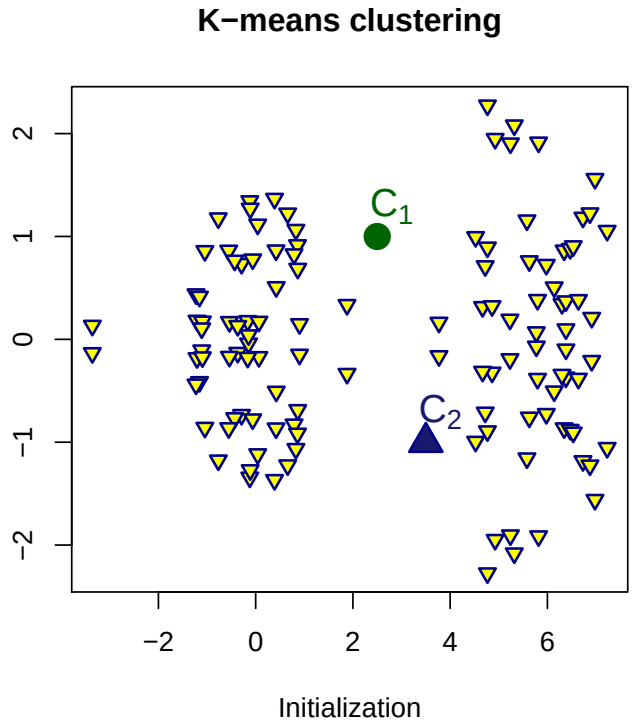
An illustrative example :



K-means clustering

Alternating minimization

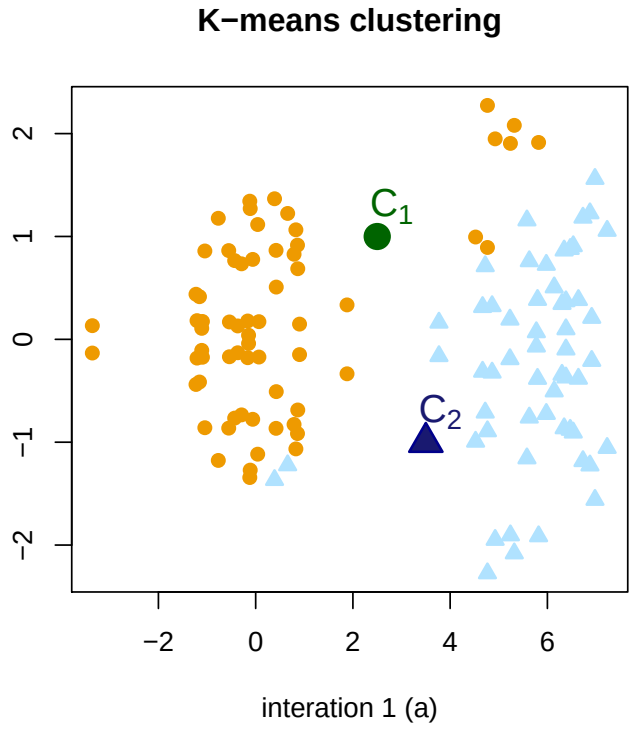
An illustrative example :



K-means clustering

Alternating minimization

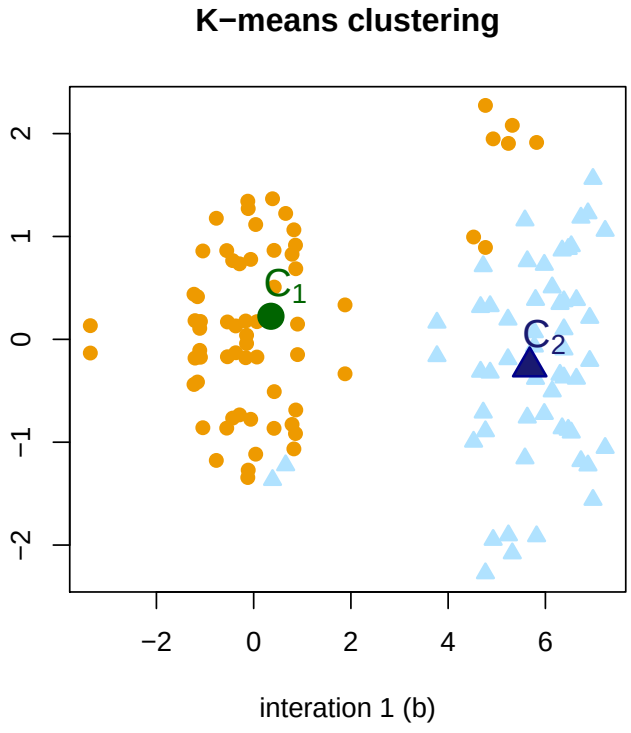
An illustrative example :



K-means clustering

Alternating minimization

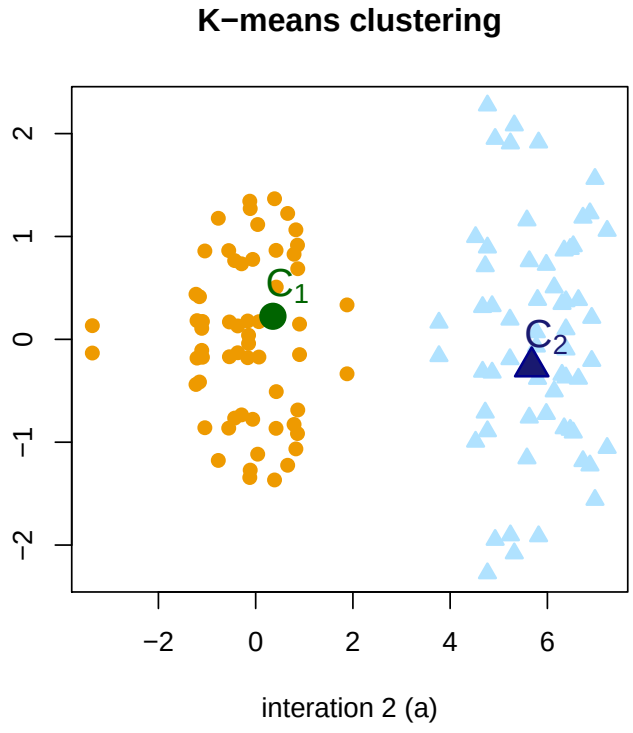
An illustrative example :



K-means clustering

Alternating minimization

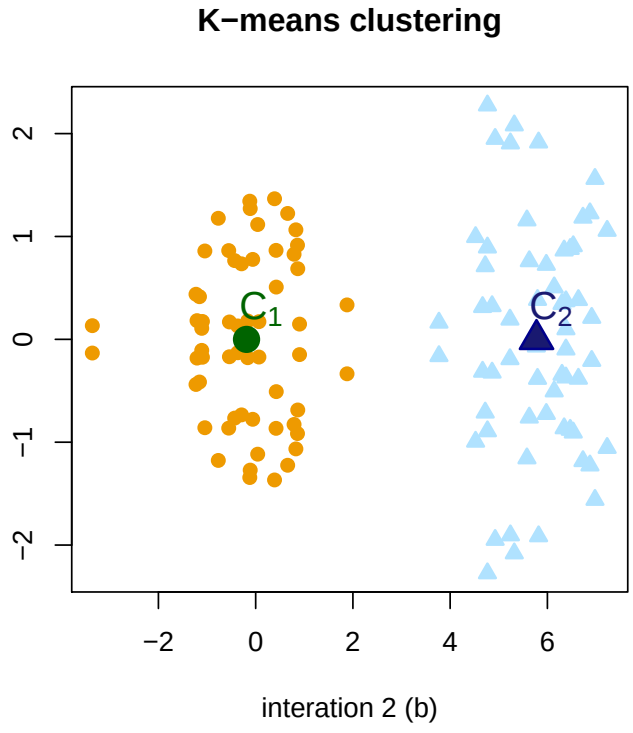
An illustrative example :



K-means clustering

Alternating minimization

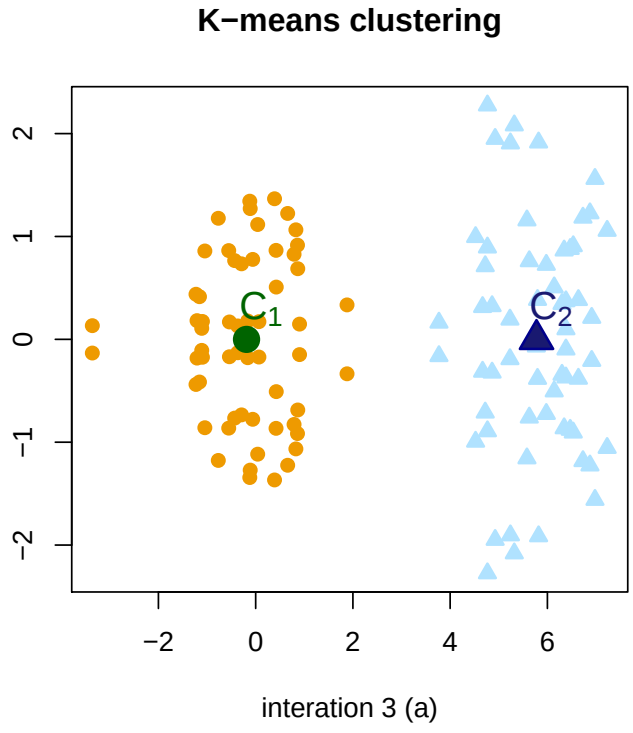
An illustrative example :



K-means clustering

Alternating minimization

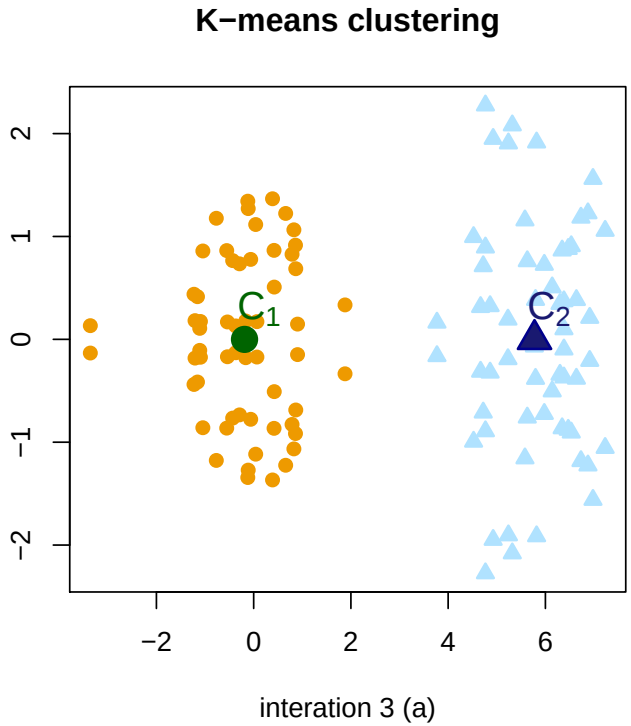
An illustrative example :



K-means clustering

Alternating minimization

An illustrative example :

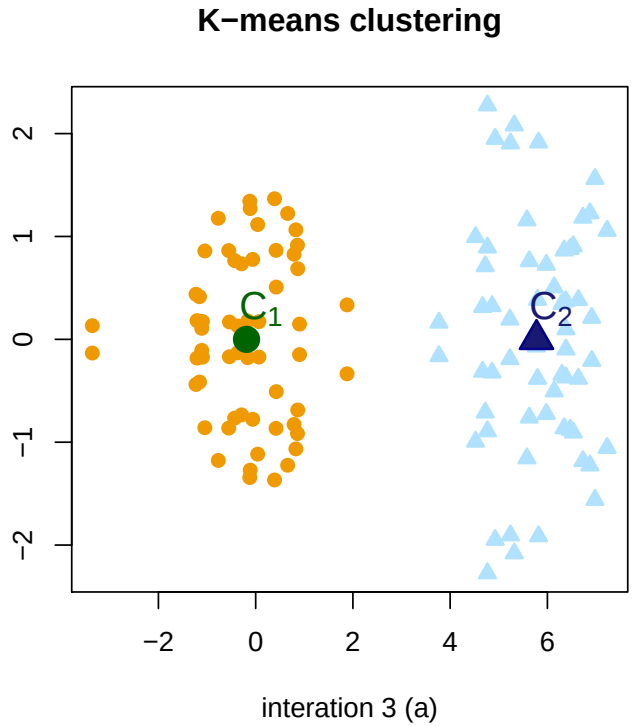


In general, a dozen iterations are sufficient for the algorithm to converge ...

K-means clustering

Alternating minimization

An illustrative example :

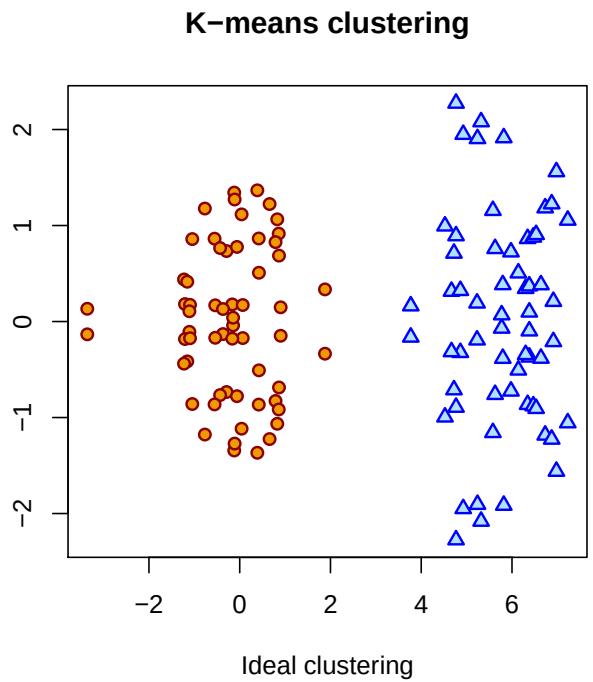


In general, a dozen iterations are sufficient for the algorithm to converge to a local minimum.

K-means clustering

Alternating minimization

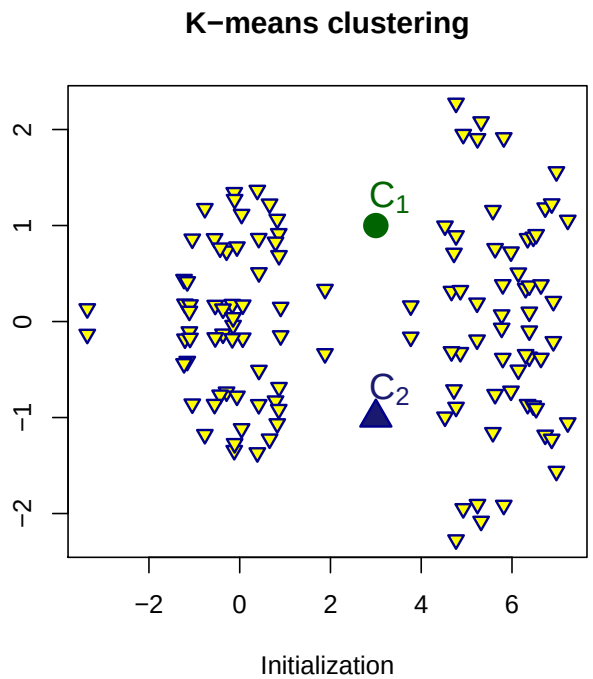
Not all initializations are good :



K-means clustering

Alternating minimization

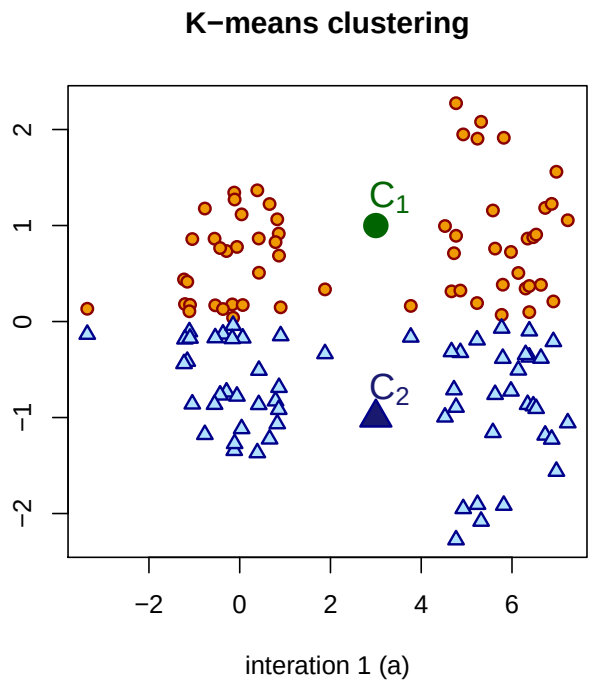
Not all initializations are good :



K-means clustering

Alternating minimization

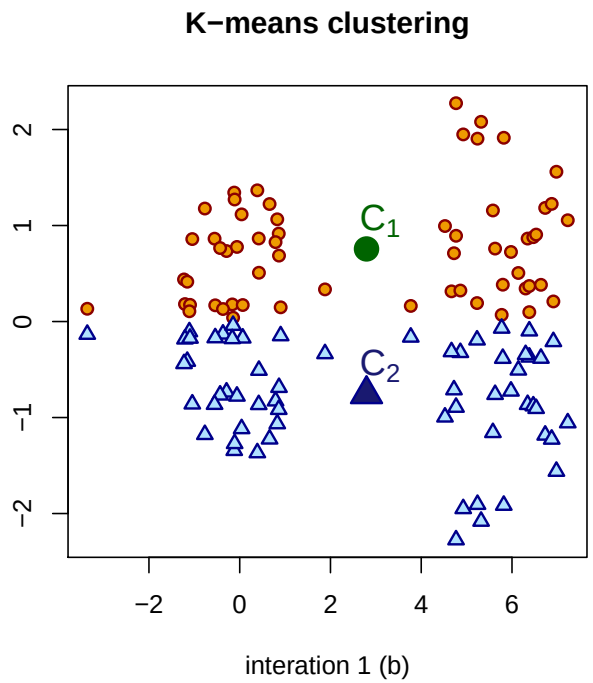
Not all initializations are good :



K-means clustering

Alternating minimization

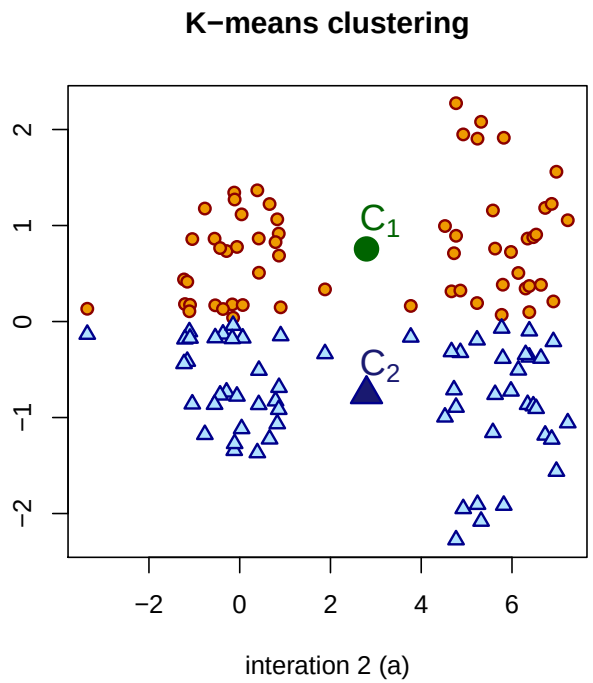
Not all initializations are good :



K-means clustering

Alternating minimization

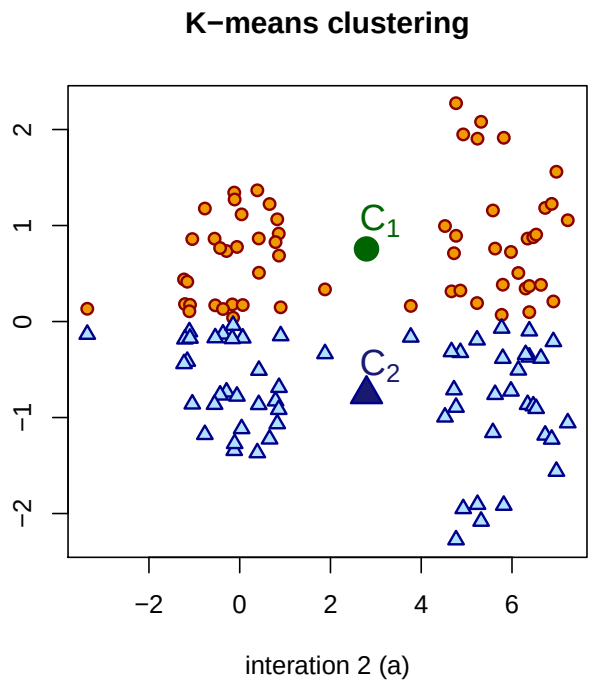
Not all initializations are good :



K-means clustering

Alternating minimization

Not all initializations are good :

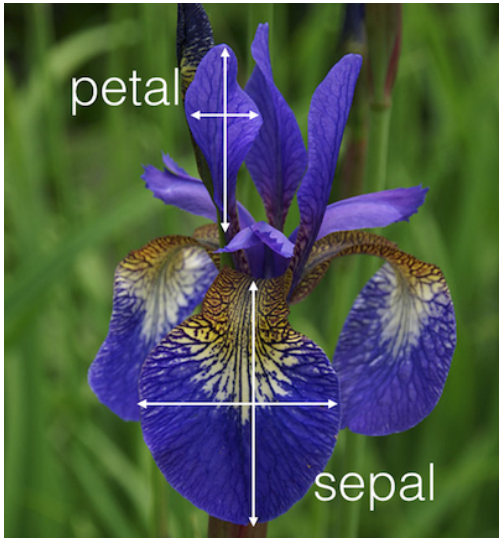


Most implementations (and this is the case for the R function `kmeans`) compute the clusterings for several initializations and select the one having the minimal cost.

K-means clustering

The iris dataset

The Iris flower data set was introduced by Ronald Fisher in his 1936 paper. It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.

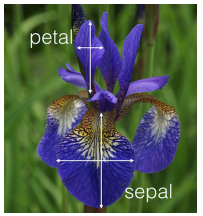


Fisher's *Iris* Data

Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*

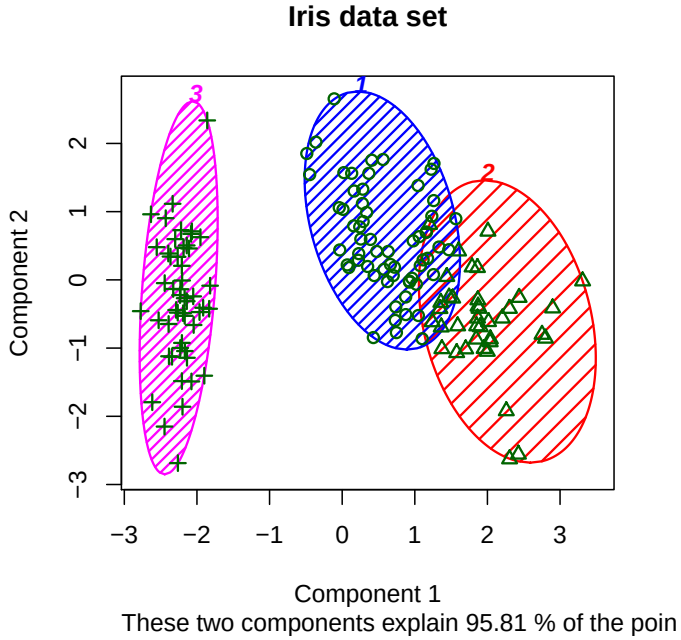
K-means clustering

R code



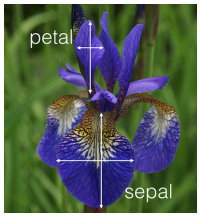
Here is a toy example of clustering with **R**.

```
> Data <- iris[,1:4]
> clus <- kmeans(Data,3)
> clus$size
[1] 38 50 62
> title <- "Iris data set"
> clusplot(Data, clus$cluster,
+       color=TRUE, shade=TRUE,
+       labels=4, lines=0,
+       main=title)
```



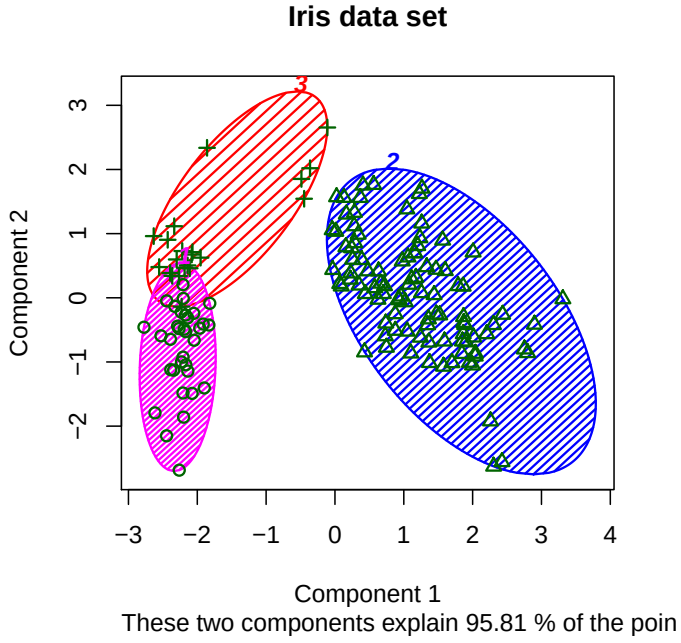
K-means clustering

R code



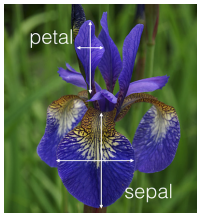
Here is a toy example of clustering with **R**.

```
> Data <- iris[,1:4]
> clus <- kmeans(Data,3)
> clus$size
[1] 33 96 21
> title <- "Iris data set"
> clusplot(Data, clus$cluster,
+       color=TRUE, shade=TRUE,
+       labels=4, lines=0,
+       main=title)
```



K-means clustering

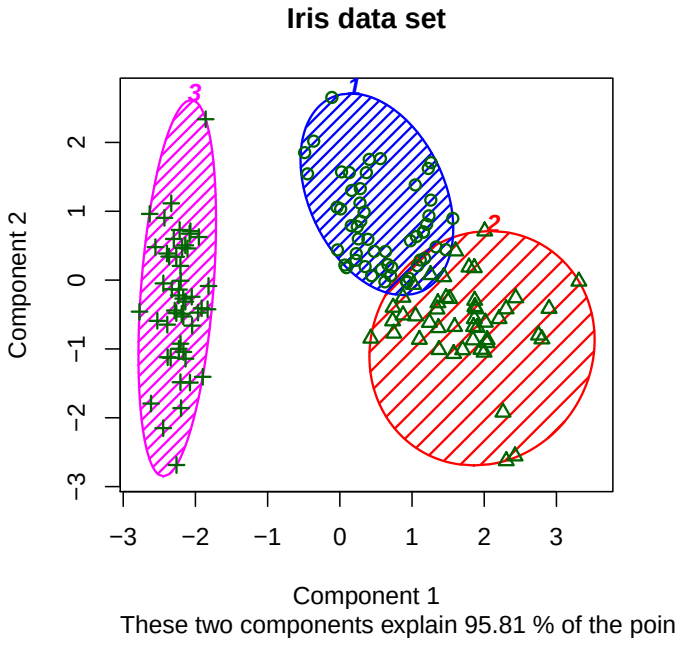
R code



Pay attention to :

- **use several initializations for reducing the randomness of the result**
- **normalize the columns of the data matrix**

```
> Data <- scale(iris[,1:4])
> clus <- kmeans(Data,3,
+                    nstart=100)
> clus$size
[1]   53 47 50
> title <- "Iris data set"
> clusplot(Data, clus$cluster,
+      color=TRUE, shade=TRUE,
+      labels=4, lines=0,
+      main=title)
```



K-means clustering

Remarks

- Breaking the ties In the step of assigning data points to clusters it may happen that two centers are the distance from a data point. Such a tie is usually broken at random using a coin toss.
- K-medians If the data is likely to contain outliers, it might be better to replace the squared Euclidean distance by the “manhattan” distance :

$$\|\mathbf{X}_i - \mathbf{X}_j\|_1 = \sum_{\ell=1}^{p} |x_{i\ell} - x_{j\ell}|$$

and to minimize the cost function

$$\sum_{k=1}^{K} \sum_{i \in G_k} \|\mathbf{X}_i - \mathbf{C}_k\|_1 .$$

This method is referred to as k-medians since $\mathbf{C}_k$ is necessarily the median of the cluster $\{\mathbf{X}_i : i \in G_k\}$.

- PAM More generally, one can consider any measure of dissimilarity between data points. The resulting algorithm is called partitioning around medoids.

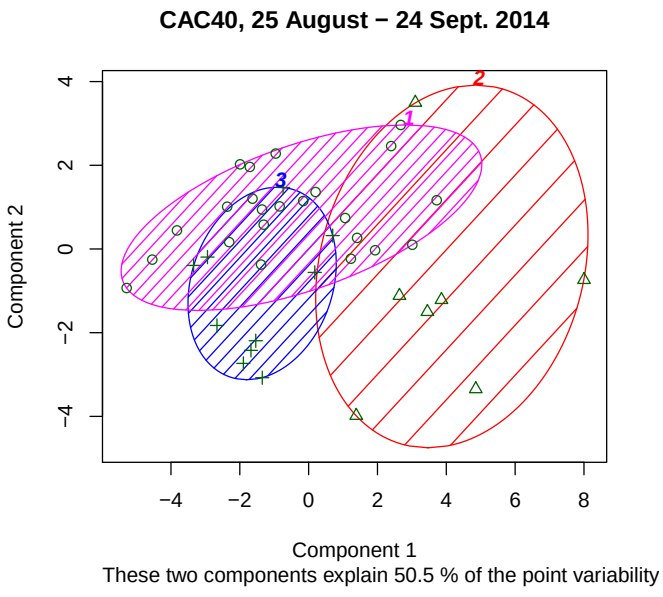
K-means clustering

Clustering CAC40 companies

Let us consider the problem of clustering the CAC40 companies according to their stock prices during the past 2 months.

- Each company is described by its daily stock returns (44 real values), the intraday variation (45 positive values) and the daily volume of exchanges (45 positive values). Data is downloaded from http://www.abcbourse.com/.
- We first determine the number of clusters using the function kmeansruns

```
> data <- scale(CAC40, center =FALSE)
> clus <- kmeansruns(data,3:10)
> k <-clus$bestk
> print(k)
[1] 3
> clus <- kmeans(data, k, nstart=100)
> clusplot(data, clus$cluster,
+      color=TRUE, shade=TRUE,
+      labels=4, lines=0,
+      main=title)
```



Clustering CAC40 companies

The result of k-means

Cluster 1 Size :23	Cluster 2 Size :7	Cluster 3 Size :10
Total	Credit Agr	Accor
L'oreal	Alcatel-Lucent	Bouygues
Lvmh	Societe Generale	Lafarge
Schneider El	Bnp Paribas	Michelin
Veolia Env	Renault	Saint Gobain
GDF Suez	Orange	Vinci
Alstom	Arcelor Mittal	Valeo
EDF		Publicis Groupe
Pernod Ric		Technip
Danone		Gemalto
...		

ENSAE ParisTech

Expectation maximization for Gaussian mixtures

Gaussian Mixture (GM) Model based clustering

Theoretical background

- The data points $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are assumed to be generated at random according to the following mechanism :
  - n "lablels" $Z_1, \ldots, Z_n$ are independently generated according to a distribution π on the set $\{1, \ldots, K\}$ (if $Z_i = k$ then $\mathbf{X}_i$ belongs to the k^{th} cluster).
  - for each $i = 1, \ldots, n$, if $Z_i = k$, then the data point $\mathbf{X}_i$ is drawn from a multivariate Gaussian distribution with a mean $\boldsymbol{\mu}_k$ and a covariance matrix Σ_k.
- Only $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are observed. The goal is to find labels $\hat{Z}_1, \ldots, \hat{Z}_n$ such that the probability of the event $\{\hat{Z}_i = Z_i, \forall i\}$ is as high as possible.

Gaussian Mixture (GM) Model based clustering

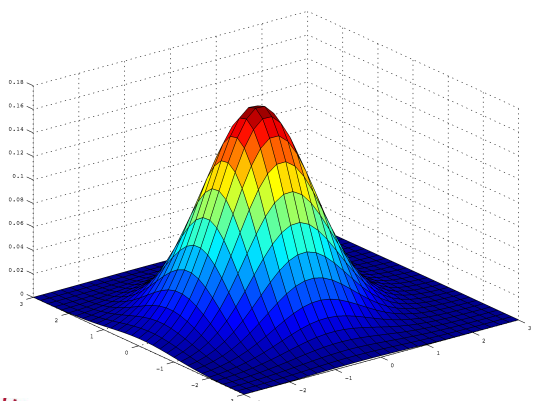
From a mathematical point of view, $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and have the density :

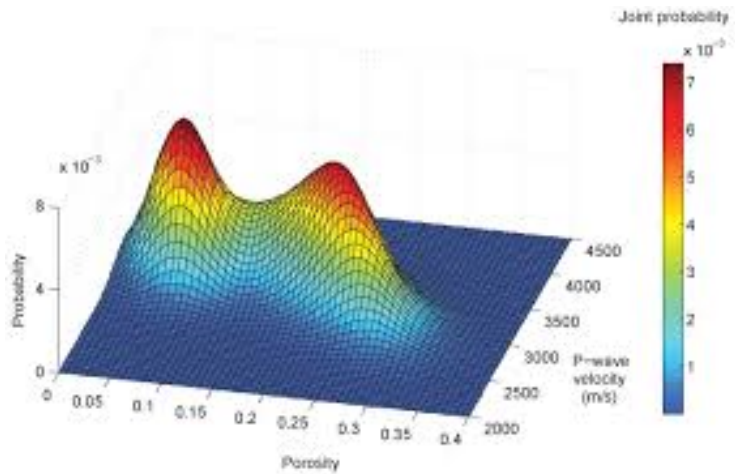
$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

with $\varphi(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ being the Gaussian density

$$\varphi(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}\|\Sigma^{-1/2}(\boldsymbol{x} - \boldsymbol{\mu})\|^2 \right\}.$$

Gaussian density in 2D	Density of a GM in 2D ($K = 2$)





ENSAE ParisTech

Gaussian Mixture (GM) Model based clustering

From clustering to estimation

- Since $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and have the density :

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

  if the parameters $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ of the model were known, we would determine the cluster number assigned to $\boldsymbol{x}$ by maximizing w.r.t. k

$$\mathbf{P}\big(Z = k|\mathbf{X} = \boldsymbol{x}\big) = \frac{\pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{p(\boldsymbol{x})}.$$

Gaussian Mixture (GM) Model based clustering

From clustering to estimation

- Since $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and have the density :

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

  if the parameters $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ of the model were known, we would determine the cluster number assigned to $\boldsymbol{x}$ by maximizing w.r.t. k

$$\mathbf{P}\big(Z = k|\mathbf{X} = \boldsymbol{x}\big) = \frac{\pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{p(\boldsymbol{x})}.$$

- Therefore, it is natural to first estimate $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ by $\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k)_k, (\hat{\Sigma}_k)_k\}$ and then to set

$$\hat{Z} = \arg \max_{k=1,\ldots,K} \hat{\pi}_k \varphi(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k).$$

Gaussian Mixture (GM) Model based clustering

From clustering to estimation

- Since $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and have the density :

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$$

  if the parameters $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ of the model were known, we would determine the cluster number assigned to $\boldsymbol{x}$ by maximizing w.r.t. k

$$\mathbf{P}\big(Z = k|\mathbf{X} = \boldsymbol{x}\big) = \frac{\pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)}{p(\boldsymbol{x})}.$$

- Therefore, it is natural to first estimate $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ by $\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k)_k, (\hat{\Sigma}_k)_k\}$ and then to set

$$\hat{Z} = \arg \max_{k=1,\ldots,K} \hat{\pi}_k \varphi(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_k, \hat{\Sigma}_k).$$

- **Problem :** how to estimate these parameters ?

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Estimation problem

Estimate the quantities $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ based on the observation of n independent random variables with density $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$.

- **Main difficulty :** the maximum likelihood estimator

$$\left\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k)_k, (\hat{\Sigma}_k)_k\right\}^{ML} = \arg \max_{\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}} \sum_{i=1}^{n} \log p(\mathbf{X}_i)$$

  is, in practice, impossible to compute.

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Estimation problem

Estimate the quantities $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ based on the observation of n independent random variables with density $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$.

- **Main difficulty :** the maximum likelihood estimator

$$\left\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k)_k, (\hat{\Sigma}_k)_k\right\}^{ML} = \arg \max_{\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \varphi(\mathbf{X}_i|\boldsymbol{\mu}_k, \Sigma_k) \right\}$$

  is, in practice, impossible to compute.

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Estimation problem

Estimate the quantities $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ based on the observation of n independent random variables with density $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$.

- **Main difficulty :** the maximum likelihood estimator

$$\left\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k)_k, (\hat{\Sigma}_k)_k\right\}^{ML} = \arg \max_{\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \varphi(\mathbf{X}_i|\boldsymbol{\mu}_k, \Sigma_k) \right\}$$

  is, in practice, impossible to compute.

- **Work-around :** use the identity

$$\log \left\{ \sum_{k=1}^{K} \pi_k a_k \right\} = \max_{\boldsymbol{\nu}} \sum_{k=1}^{K} \left(\nu_k \log(\pi_k a_k) - \nu_k \log \nu_k \right).$$

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Estimation problem

Estimate the quantities $\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}$ based on the observation of n independent random variables with density $p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \varphi(\boldsymbol{x}|\boldsymbol{\mu}_k, \Sigma_k)$.

- **Main difficulty :** the maximum likelihood estimator

$$\left\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k)_k, (\hat{\Sigma}_k)_k\right\}^{ML} = \arg \max_{\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k)_k, (\Sigma_k)_k\}} \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \varphi(\mathbf{X}_i|\boldsymbol{\mu}_k, \Sigma_k) \right\}$$

  is, in practice, impossible to compute.

- **Work-around :** use the identity

$$\log \left\{ \sum_{k=1}^{K} \pi_k a_k \right\} = \max_{\substack{\boldsymbol{\nu} \in \mathbb{R}_+^K \\ \sum \nu_k = 1}} \sum_{k=1}^{K} \left(\nu_k \log a_k - \nu_k \log(\nu_k/\pi_k) \right).$$

ENSAE ParisTech

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Let us denote $\Omega = \left\{\boldsymbol{\pi}, (\boldsymbol{\mu}_k), (\Sigma_k)\right\}$ and $\hat{\Omega}^{ML} = \left\{\hat{\boldsymbol{\pi}}, (\hat{\boldsymbol{\mu}}_k), (\hat{\Sigma}_k)\right\}^{ML}$.

Using the previous formulae, we can write $\hat{\Omega}^{ML}$ as follows

$$\hat{\Omega}^{ML} = \arg\max_{\Omega} \max_{\boldsymbol{\nu}_i} \underbrace{\sum_{i=1}^{n}\sum_{k=1}^{K}\left\{\nu_{ik}\log\left(\pi_k\varphi(\mathbf{X}_i|\boldsymbol{\mu}_k,\Sigma_k)\right) - \nu_{ik}\log\nu_{ik}\right\}}_{\boldsymbol{G}(\Omega,\boldsymbol{\nu})}. \tag{1}$$

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Let us denote $\Omega = \left\{\pi, (\mu_k), (\Sigma_k)\right\}$ and $\hat{\Omega}^{ML} = \left\{\hat{\pi}, (\hat{\mu}_k), (\hat{\Sigma}_k)\right\}^{ML}$.

Using the previous formulae, we can write $\hat{\Omega}^{ML}$ as follows

$$\hat{\Omega}^{ML} = \arg \max_{\Omega} \max_{\nu} \boldsymbol{G}(\Omega, \nu). \tag{1}$$

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Let us denote $\Omega = \left\{\pi, (\mu_k), (\Sigma_k)\right\}$ and $\hat{\Omega}^{ML} = \left\{\hat{\pi}, (\hat{\mu}_k), (\hat{\Sigma}_k)\right\}^{ML}$.

Using the previous formulae, we can write $\hat{\Omega}^{ML}$ as follows

$$\hat{\Omega}^{ML} = \arg \max_{\Omega} \max_{\nu} \boldsymbol{G}(\Omega, \nu). \tag{1}$$

- For a fixed Ω, the maximum w.r.t. ν in (1) is explicitly computable.
- For a fixed ν, the maximum w.r.t. Ω in (1) is explicitly computable as well.

Gaussian Mixture (GM) Model based clustering

EM algorithm : main idea

Let us denote $\Omega = \left\{\pi, (\mu_k), (\Sigma_k)\right\}$ and $\hat{\Omega}^{ML} = \left\{\hat{\pi}, (\hat{\mu}_k), (\hat{\Sigma}_k)\right\}^{ML}$.

Using the previous formulae, we can write $\hat{\Omega}^{ML}$ as follows

$$\hat{\Omega}^{ML} = \arg \max_{\Omega} \max_{\nu} \boldsymbol{G}(\Omega, \nu). \tag{1}$$

- For a fixed Ω, the maximum w.r.t. ν in (1) is explicitly computable.
- For a fixed ν, the maximum w.r.t. Ω in (1) is explicitly computable as well.

EM-algorithm

Initialize Ω. Then iteratively (until convergence)
- update ν by solving (1) with fixed Ω
- update Ω by solving (1) with fixed ν.

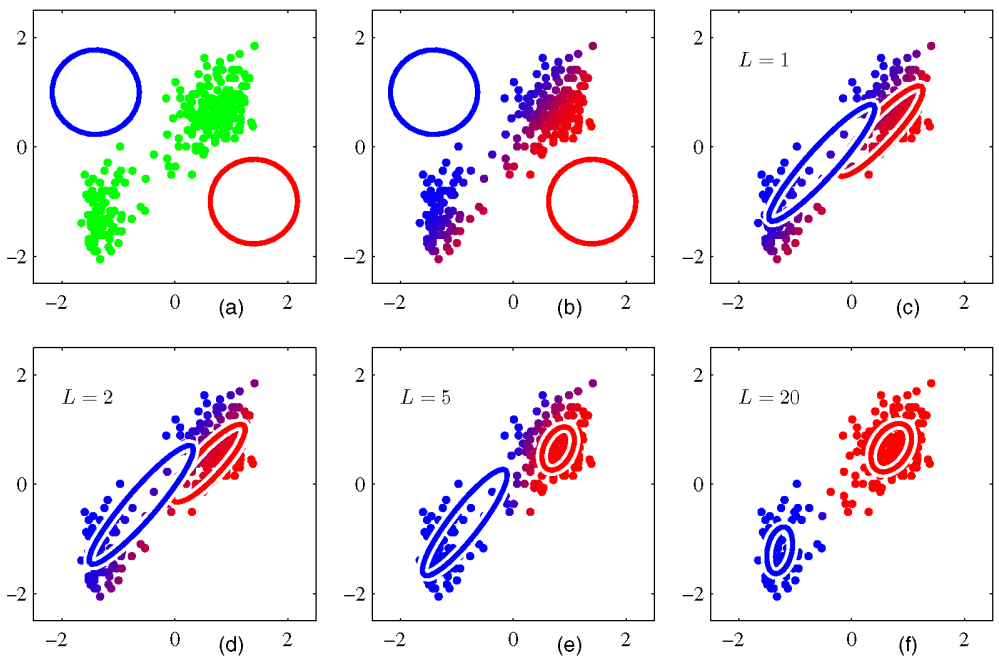
Gaussian Mixture (GM) Model based clustering

EM algorithm : how it works

EM-algorithm

Initialize Ω. Then iteratively (until convergence)

- update ν by solving (1) with fixed Ω
- update Ω by solving (1) with fixed ν.



Gaussian Mixture (GM) Model based clustering

EM algorithm : summary and remarks

- The goal of the EM-algorithm is to approximate the maximum likelihood estimator.
- The EM-algorithm solves a nonconvex optimization problem. Therefore, there is no guarantee that it finds the global optimum. Generally, it does not !
- To apply the EM algorithm, the sample size n should be significantly larger than $p \times k + p^2$.
- One can adapt the EM algorithm to other (non Gaussian) distributions.
- It is possible in the EM algorithm to assume that all the Σ_k's are equal, or that they are all diagonal.
- The auxiliary values ν_{ik} computed by the EM-algorithm estimate the probability of $\mathbf{X}_i$ to belong to the cluster k.

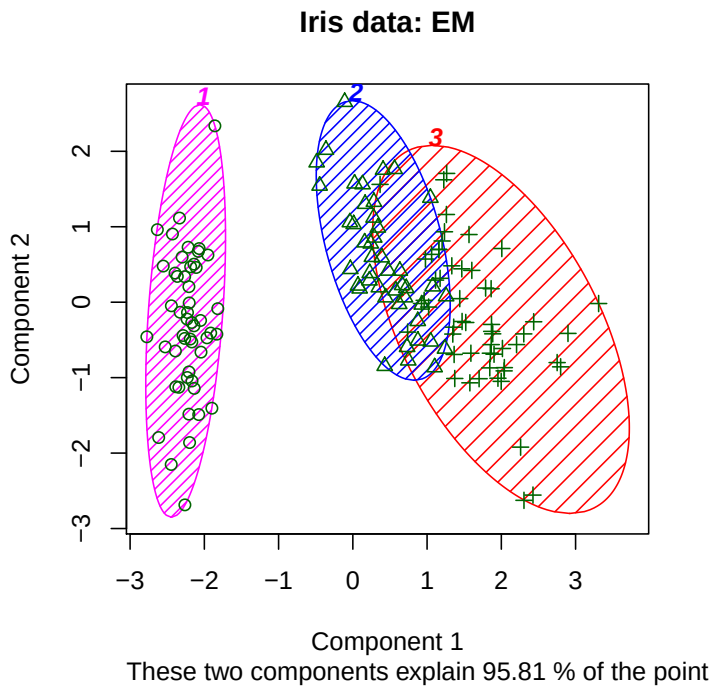
ENSAE ParisTech

Gaussian Mixture (GM) Model based clustering

EM algorithm : R code

The EM-algorithm is implemented in the R-package `mclust`.

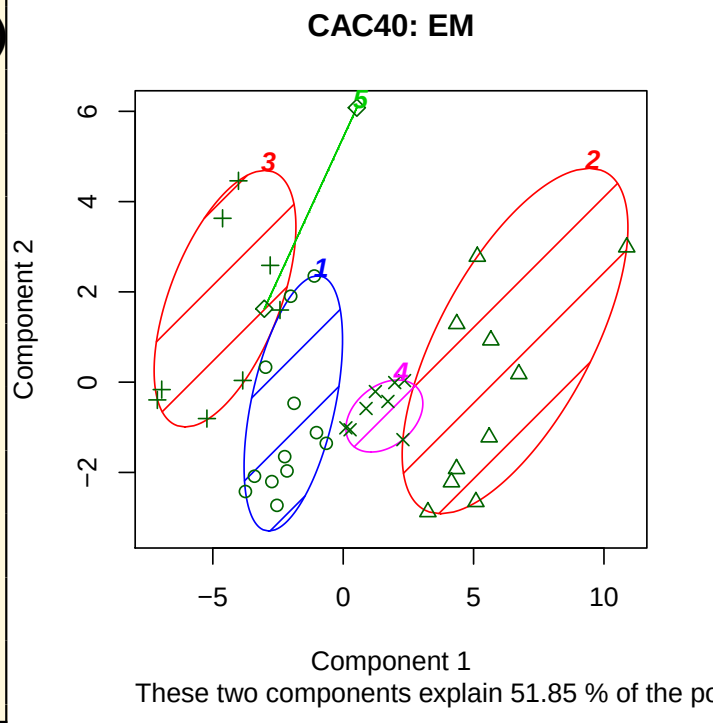
```
> data <- (iris[,1:4])
> irisC <- Mclust(data, G=3)
> clusters <- irisC$classification
> table(clusters)
clusters
 1  2  3
50 45 55
> clusplot(data, clusters,
+      color=TRUE, shade=TRUE,
+      labels=4, lines=0,
+      main="Iris data: EM")
```



Gaussian Mixture (GM) Model based clustering

R code for CAC 40 dataset

```
> data <- scale(CAC40, center =FALSE)
> CAC.cl <- Mclust(data)
> clusters <- CAC.cl$classification
> table(clusters)
clusters
 1  2  3  4  5
12 10  8  8  2
> clusplot(data, clusters,
+      color=TRUE, shade=TRUE,
+      labels=4, lines=0,
+      main="CAC40: EM")
```



Clustering CAC40 companies

The result of EM

Cluster 1 Size :12	Cluster 2 Size :10	Cluster 3 Size :8	Cluster 4 Size :8	Cluster 5 Size :2
Safran				
Air Liq	Credit Agr			
L'oreal	Total	Solvay	Carrefour	
Accor	Axa	Michelin	Sanofi	
Bouygues	Vivendi	Kering	Danone	
Pernod Ric	Alcatel-Luc.	Unibail-Rod.	Schneider El	Lafarge
Lvmh	Soc. Gen.	Valeo	Veolia Env	Renault
Essilor Intl	Bnp Paribas	Publicis Gr.	Saint Gobain	
Cap Gemini	Orange	Technip	Vinci	
Alstom	GDF Suez	Gemalto	Airbus Group	
EDF	Arcelor Mit.			
Legrand SA				

ENSAE ParisTech


рахмат
danke
謝謝
ngiyabonga
Баярлалаа
teşekkür ederim
спасибо
mersi
barka
welalin
tack
spas
kia ora
faafetai lava
vinaka
dank je
misaotra
paldies
grazzi
mahalo
tapadh leat
kiitos
dankie
спасибі
blagodaram
matondo
gracias
хвала
asante
manana
dhanyavad
thank you
obrigada
murakoze
tenki
enkosi
bedankt
hvala
maururu
köszönöm
djiere dieuf
tau
mochchakkeram
дякую
mamnun
chokrane
dziękuję
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
sobodi
dekuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
najis tuke
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
merce
мерси
kam sah hamnida
rahmat
감사합니다
xièxie
ευχαριστώ
merci