

# Unsupervised learning

## Clustering

Statistical methods for data analysis – Machine learning

Alberto Castellini  
University of Verona

- **Supervised methods** (learning with a teacher):
  - **Input:** predictor variables  $X^T = (X_1, \dots, X_p)$
  - **Output:**  $Y$
  - Predictions based on the **training set**  $(x_1, y_1), \dots, (x_N, y_N)$  of previously solved cases
  - **Loss function**  $L(y, \hat{y})$ , such as  $L(y, \hat{y}) = (y - \hat{y})^2$
  - Supposing  $(X, Y)$  random variables supervised learning is a **density estimation problem**:

**Determining the properties of the conditional density  $\Pr(Y|X)$**

**E.g.: location parameters that minimize the expected error**

$$\mu(x) = \operatorname{argmin}_{\theta} E_{Y|X} L(Y, \theta).$$

- **Unsupervised methods** (learning **without** a teacher):
  - **Input:** predictor variables  $X^T = (X_1, \dots, X_p)$
  - **Output:** not available
  - **Goal:** infer the **properties of the joint probability  $\Pr(X)$**  without the help of a supervisor/teacher
  - In **low dimensional problems ( $p \leq 3$ )**  $\Pr(X)$  can be directly estimated and graphically represented
  - In **high dimensional problems** descriptive statistics methods are used to characterize  $\Pr(X)$ 
    - **Low dimensional manifolds** representing high data density may be identified by **PCA** or other **dimensionality reduction** methods
    - **Cluster analysis** attempts to find multiple convex regions of the  $X$ -space that contain **modes** of  $\Pr(X)$
  - No direct measure of success (as loss function)

# *Unsupervised learning methods*

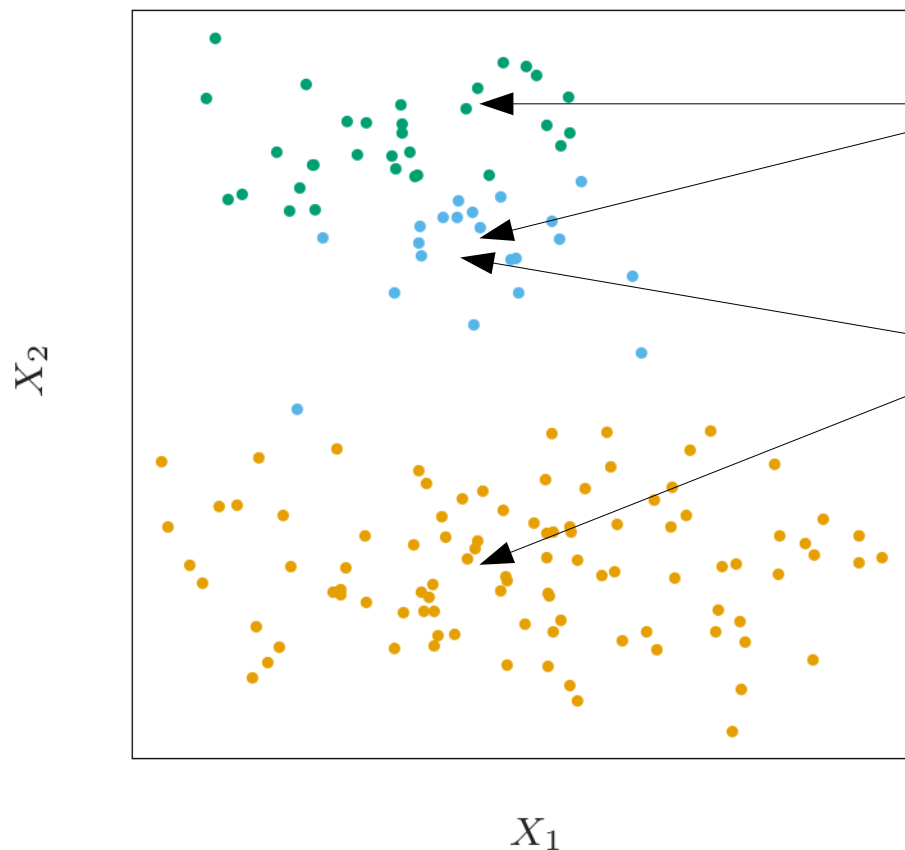
- Association rules
- Clustering analysis
  - K-means
  - K-medoids
  - Gaussian Mixture Models
  - Hierarchical clustering
- Self-organizing maps
- Principal components, curves and subspaces
  - Spectral clustering
- Matrix factorization
- ...

# Unsupervised learning methods

- Association rules
- Clustering analysis
  - K-means
  - K-medoids
  - Gaussian Mixture Models
  - Hierarchical clustering
- Self-organizing maps
- Principal components, curves and subspaces
  - Spectral clustering
- Matrix factorization
- ...

# *Cluster analysis*

**Grouping** a collection of **objects** into **subsets (clusters)** such that objects **within** each clusters are **more closely related** to one another **than** objects assigned to **different** clusters



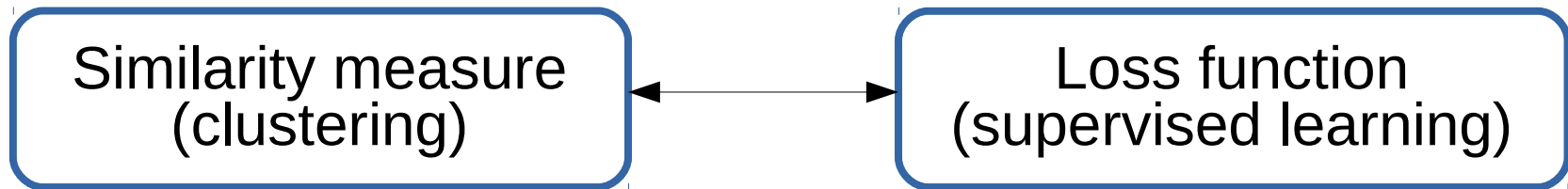
Not well separated  
(similar properties)

Why?

Well separated  
(different properties)

# Measures of similarity/dissimilarity

- Central to clustering analysis is the notion of **similarity/dissimilarity between individual objects**
- **Clustering methods** attempt to **group** the objects according to the definition of **similarity** supplied to it.



- **Examples of similarity/dissimilarity measures:**

- Euclidean distance
- Manhattan distance
- Mahalanobis distance
- Correlation
- Jaccard distance

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_{i=1}^n (q_i - p_i)^2}$$
$$d_1(\mathbf{p}, \mathbf{q}) = \|\mathbf{p} - \mathbf{q}\|_1 = \sum_{i=1}^n |p_i - q_i|$$

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}$$

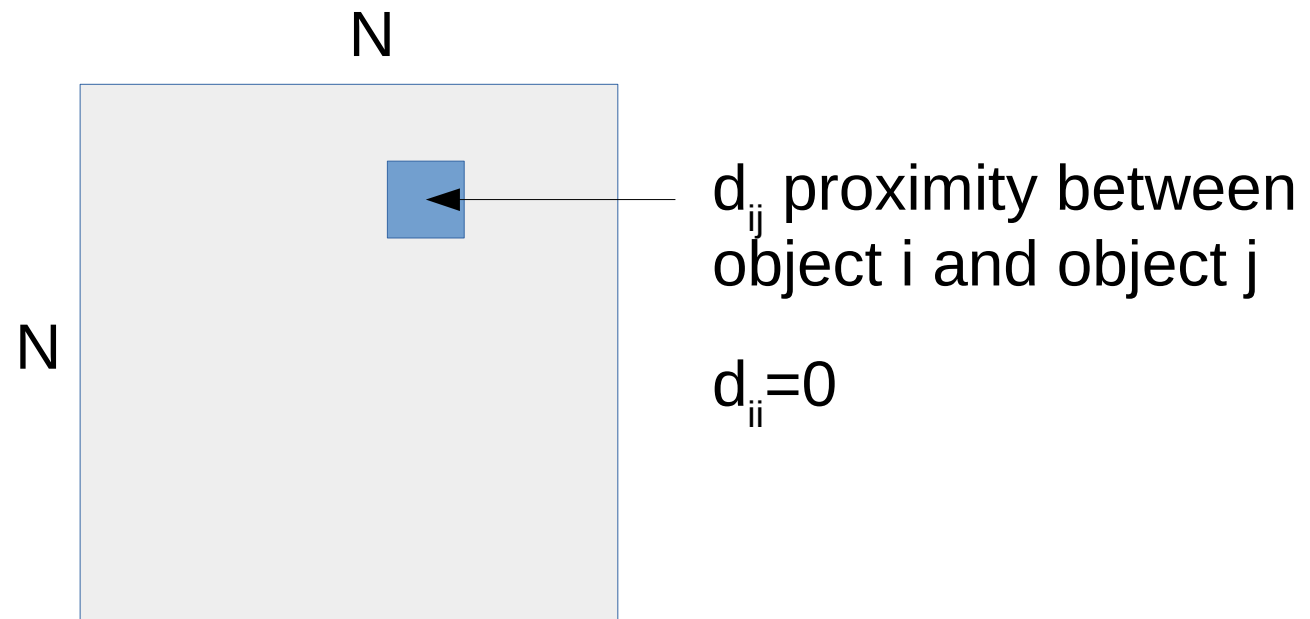
$$\rho_{X,Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



# Proximity matrices

- Sometimes the data is represented directly in terms of proximity between pairs of objects (similarities or dissimilarities).
- N x N matrix



- Dissimilarities are *distances* in the strict sense only if the triangle inequality  $d_{ii'} \leq d_{ik} + d_{i'k}$ , for all  $k \in \{1, \dots, N\}$  holds.

## Dissimilarities based on attributes (1/2)

- Usually we have measurements  $x_{ij}$   $i=1,\dots,N$ , on variables  $j=1,\dots,p$  (**attributes**).
- Then **pairwise dissimilarities** between **observations** can be expressed in terms of **attribute** values, that is

$$D(x_i, x_{i'}) = \sum_{j=1}^p d_j(x_{ij}, x_{i'j})$$

where  $d_j(x_{ij}, x_{i'j})$  is the dissimilarity between values of the  $j$ -th attribute.

- Most common **D** function is the **squared distance**

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

- **Other choices** are possible depending on attribute types

- **Quantitative variables**

- Absolute difference  $d(x_i, x_{i'}) = l(|x_i - x_{i'}|)$

- Correlation  $\rho(x_i, x_{i'}) = \frac{\sum_j (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_j (x_{ij} - \bar{x}_i)^2 \sum_j (x_{i'j} - \bar{x}_{i'})^2}}$

- ...

- **Ordinal variables**

- **Categorical variables**

## Object dissimilarity (1/2)

- **Dissimilarities** of  $p$  **attributes** are then **combined** into a single overall measure of **dissimilarity**  $D(x_i, x_{i'})$  **between objects**
- **Weighted average (convex combination):**

$$D(x_i, x_{i'}) = \sum_{j=1}^p w_j \cdot d_j(x_{ij}, x_{i'j}); \quad \sum_{j=1}^p w_j = 1$$

**Weight of  $j$ -th attribute**

- **Weight**  $w_j$  regulates the **relative influence of variable  $j$**  in determining the **overall dissimilarity** between objects
- All  $w_j=1$  does **NOT** give all attributes equal influence
- The relative influence of the  $j$ -th variable is  $w_j * \text{avg}(d_j)$

where

$$\bar{d}_j = \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N d_j(x_{ij}, x_{i'j})$$

## Object dissimilarity (2/2)

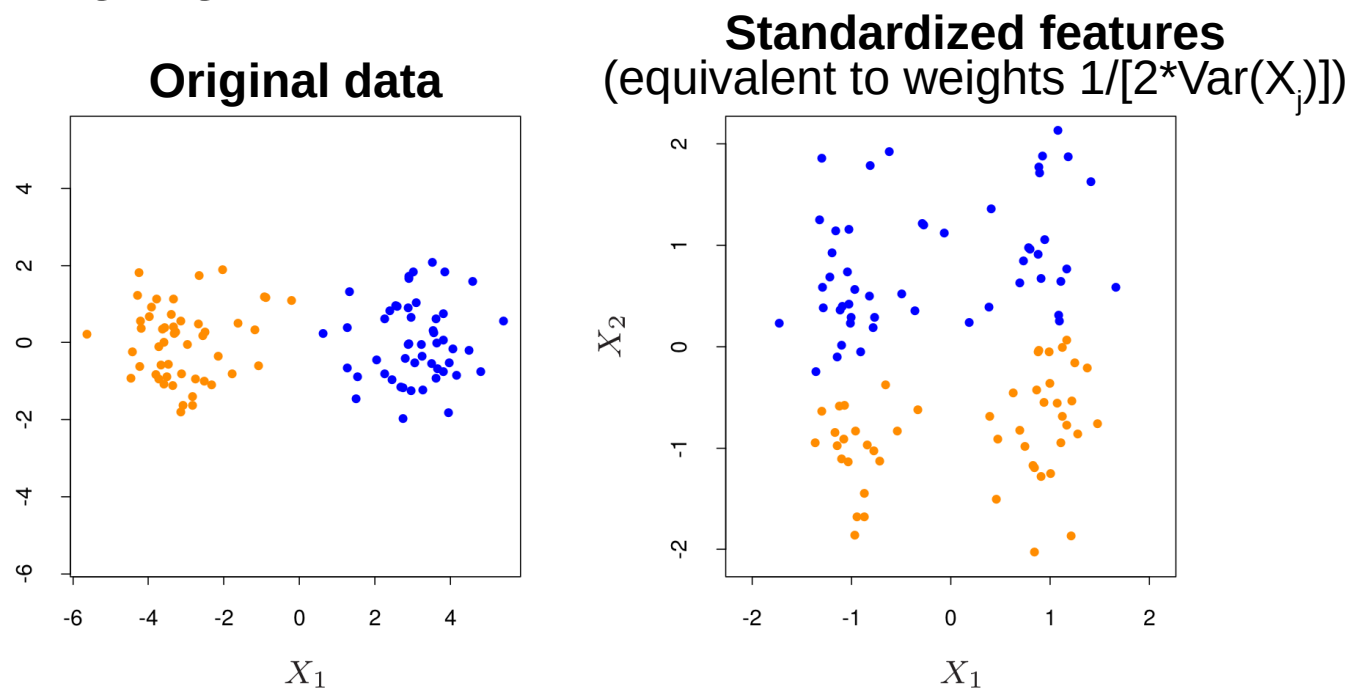
- Hence, setting  $w_j \sim 1/\text{avg}(d_j)$  gives all attributes **equal influence** on the overall dissimilarity
- This is related to **data standardization** in supervised learning
- E.g., for squared error distance

$$D_I(x_i, x_{i'}) = \sum_{j=1}^p w_j \cdot (x_{ij} - x_{i'j})^2$$
$$\bar{d}_j = \frac{1}{N^2} \sum_{i=1}^N \sum_{i'=1}^N (x_{ij} - x_{i'j})^2 = 2 \cdot \text{var}_j$$

the **relative importance** of each attribute is **proportional** to its **variance** over the data

## Attribute relative importance

- If the goal is **discovering natural grouping**, forcing equal influence among attributes can be **counterproductive**.
- **More relevant variables should have higher influence** in the object dissimilarity
- Giving all attributes equal influence tend to obscure the clusters to clustering algorithms



The choice of **appropriate dissimilarity measures** is often more **important** than the choice of the clustering algorithm

**Goal of clustering algorithms:** to partition observations into groups such that pairwise dissimilarities between observations assigned to the same cluster tend to be smaller than those in different clusters

## Algorithm types:

- Combinatorial
- Mixture modeling based
- Mode seekers

# Combinatorial algorithms

- Most popular
- No probability model
- Prespecified number of clusters  $K < N$
- Each observation labeled by an integer  $k$  in  $\{1, \dots, K\}$
- Assignments characterized by a many-to-one mapping (**encoder**):

$$k = C(i)$$

that assigns the  $i$ -th observation to the  $k$ -th cluster

Goal: seek the encoder  $C^*(i)$  that minimizes a loss (or energy) function which depends on pairwise dissimilarities

$$W(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} d(x_i, x_{i'})$$

**Within cluster point scatter**



- **Total point scatter**

$$T = \frac{1}{2} \sum_{i=1}^N \sum_{i'=1}^N d_{ii'} = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \left( \sum_{C(i')=k} d_{ii'} + \sum_{C(i') \neq k} d_{ii'} \right)$$

- **Between-cluster point scatter**

$$B(C) = \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i') \neq k} d_{ii'}$$

- **Relationships**

$$T = W(C) + B(C)$$

Minimizing  $W(C)$  is equivalent to maximizing  $B(C)$

- **Number of possible assignments** (*Jain ad Dubes, 1988*)

$$S(N, K) = \frac{1}{K!} \sum_{k=1}^K (-1)^{K-k} \binom{K}{k} k^N$$

- E.g.,  $S(10,4)=34105$ ,  $S(19,4)=10^{10}$
- Heuristic strategies: iterative greedy descent
- Initial partition
- Iterative steps for reducing the loss
- Local optima

- **Squared error distance**

$$d(x_i, x_{i'}) = \sum_{j=1}^p (x_{ij} - x_{i'j})^2 = \|x_i - x_{i'}\|^2$$

- **Within point scatter**

$$\begin{aligned} W(C) &= \frac{1}{2} \sum_{k=1}^K \sum_{C(i)=k} \sum_{C(i')=k} \|x_i - x_{i'}\|^2 \\ &= \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - \bar{x}_k\|^2, \end{aligned}$$

Euclidean distance from the centroid  
(mean vector) associated to the k-th cluster

# K-means algorithm

1) Given a cluster assignment  $C$ , the total cluster variance

$$\min_{C, \{m_k\}_1^K} \sum_{k=1}^K N_k \sum_{C(i)=k} \|x_i - m_k\|^2$$

is minimized with respect to  $\{m_1, \dots, m_K\}$  yielding the means of the currently assigned clusters.

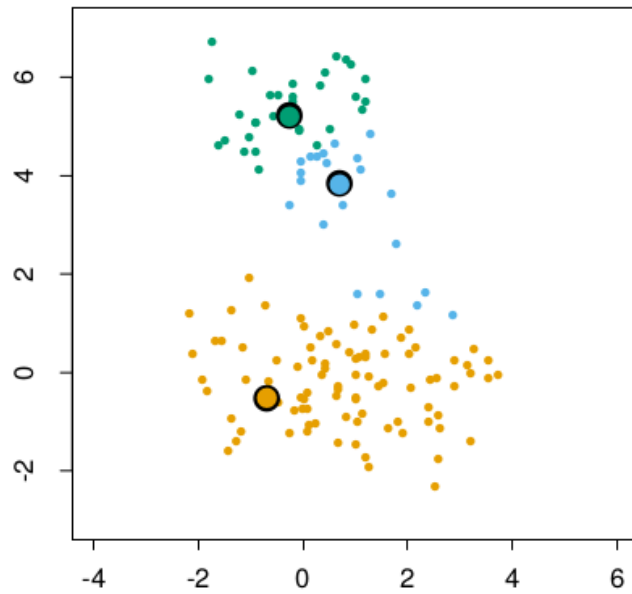
2) Given the current set of means  $\{m_1, \dots, m_K\}$  the total cluster variance is minimized by assigning each observation to the closest (current) cluster mean:

$$C(i) = \operatorname{argmin}_{1 \leq k \leq K} \|x_i - m_k\|^2$$

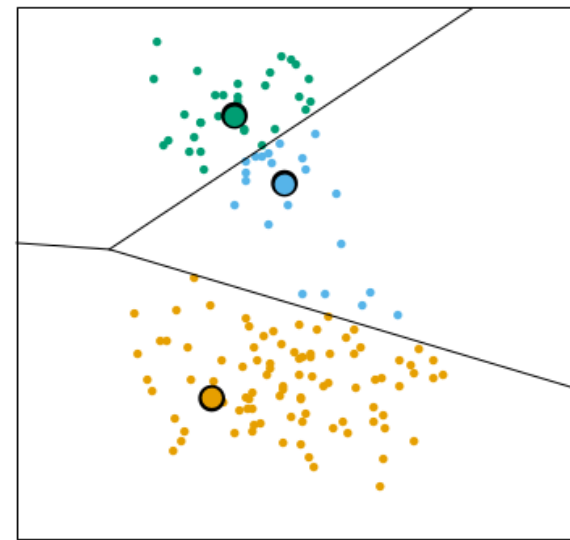
3) Iterate steps 1 and 2 until the assignments do not change.

# Successive iterations of K-means

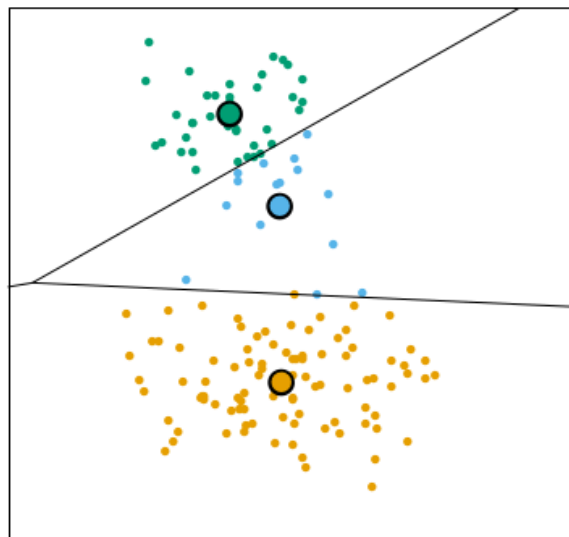
Initial Centroids



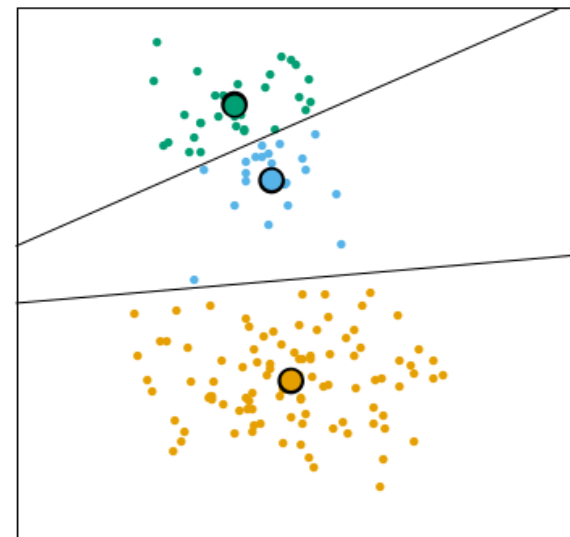
Initial Partition



Iteration Number 2



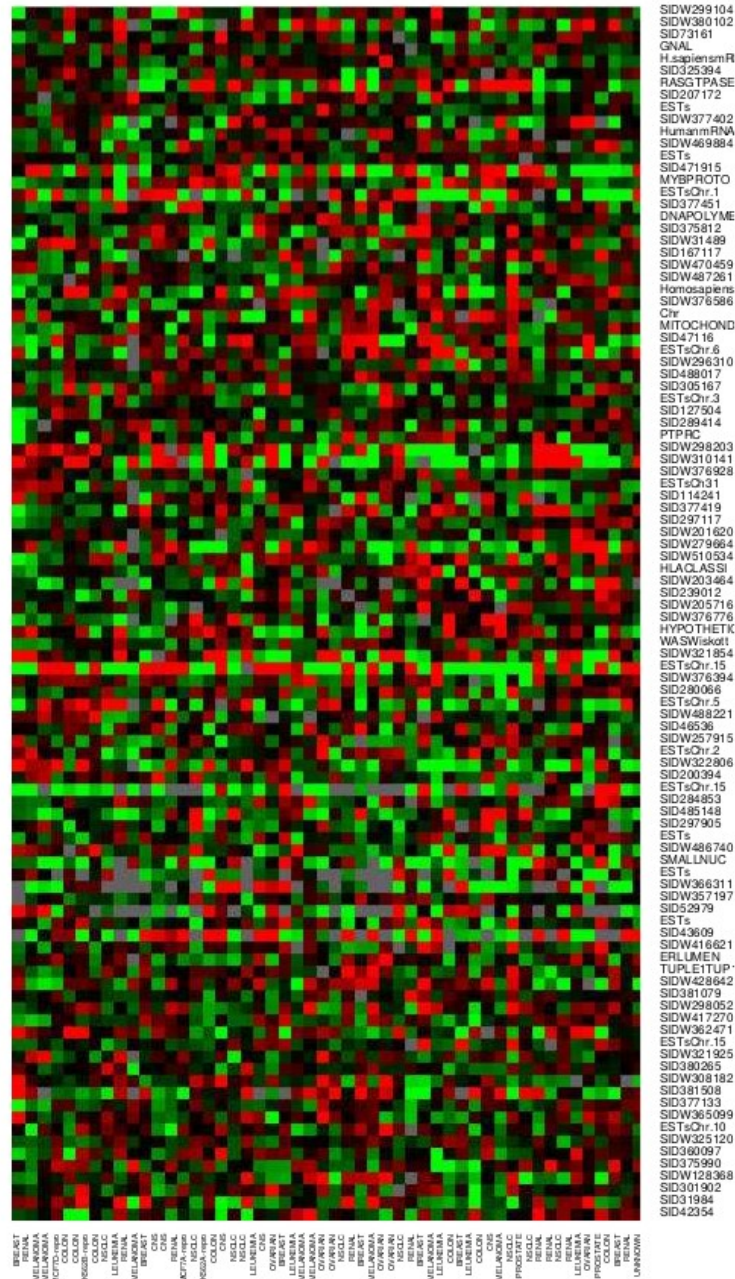
Iteration Number 20



## *Exercise: Clustering of Human tumor microarray data*

See text of Exercise 6

# Human tumor dataset



Genes

Experiments (samples)

# References

[Hastie 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition). Springer. 2009.