Machine Learning, Lecture 3: Cluster analysis I

S. Nõmm

[1]Department of Software Science, Tallinn University of Technology

Side note: notions

Given a set of data points, partition them into groups with respect to chosen similarity criteria.

- Attribute or variable is the smallest- indivisible element of the data (dataset).
- *Data point* or *observation point*- tuple of attributes or variables composed with respect to the experiment setting.
- Feature - is the set of attributes composed either with respect to some properties or by some algorithm. **NB!** The attributes of a feature may be selected on the basis of the associations between them, but the associations itself does not represent the part of the feature. Roughly speaking feature does not contain any explicit information about the relationships between its attributes.
- Pattern is defined by two sets of conditions. The first set defines the elements (features or attributes) and the second one defines associations between the attributes.
- In some cases notion of the Templates is also used.

Different approaches

- **Representative based algorithms**
- Hierarchical Clustering Algorithms
- Group-Based Statistics
- Grid- and density- based methods

Representative-Based Algorithms

- The k-Means Algorithm.
- The Kernel k-Means Algorithm
- The k-Medians Algorithm
- The k-Medoids Algorithm

K-means

The goal is to cluster the data into K clusters, whereas no labeled data are given.

- Case of unsupervised learning.
- K is the hyperparameter.

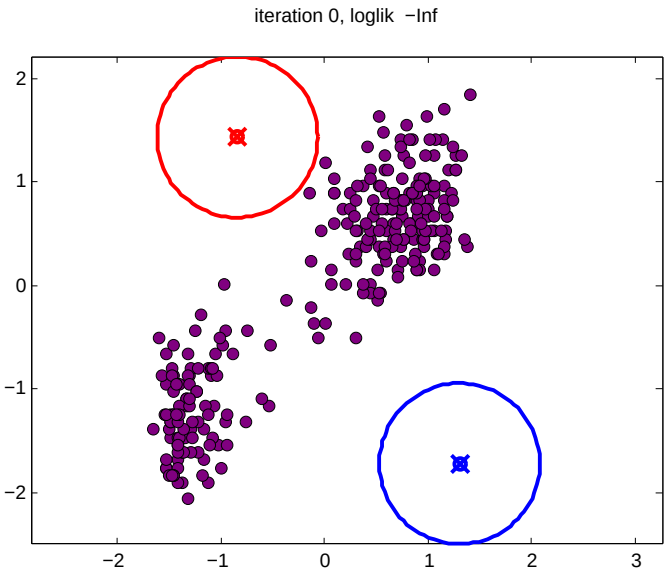
K-means clustering

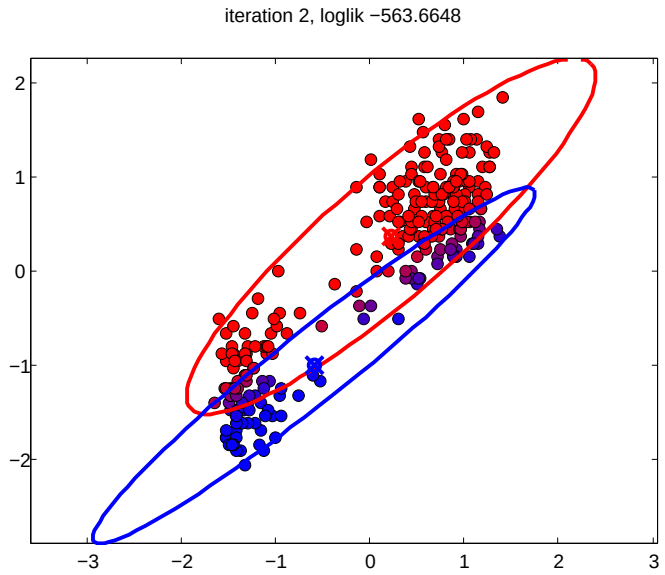
- Initialization: Generate randomly K points, called *Centroids*. Each centroid represent one of the K classes.
  repeat
  - Associate each point with the cluster represented by the closest centroid. $z_i = \arg\min_k \mid\mid x_i - \mu_k \mid\mid_2^2$. z_i - is the cluster label.
  - Update centroids for each cluster as

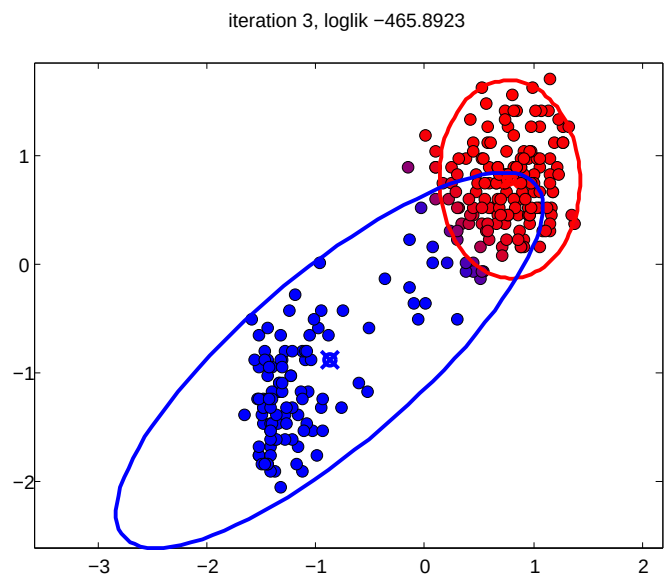
$$\mu_k = \frac{1}{N_k} \sum_{i:z_i=k} x_i$$

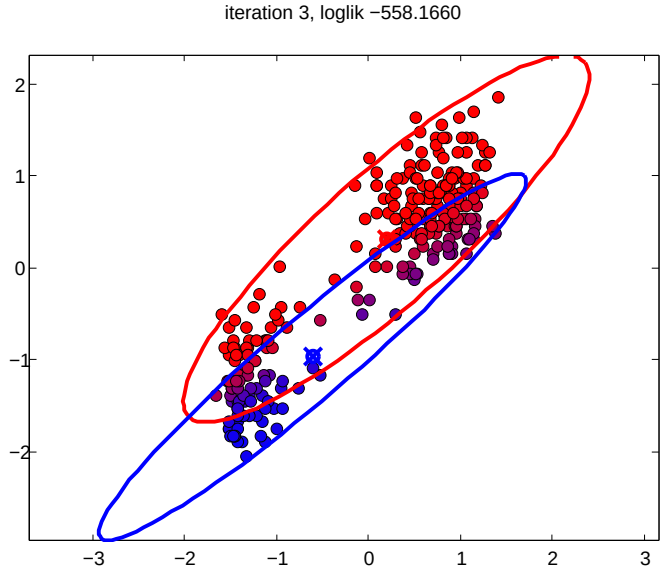
  until converged;

Example 1 of 4

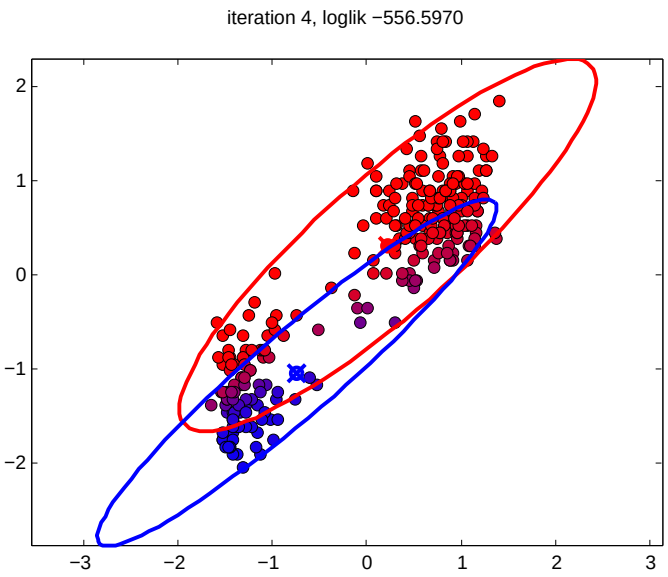


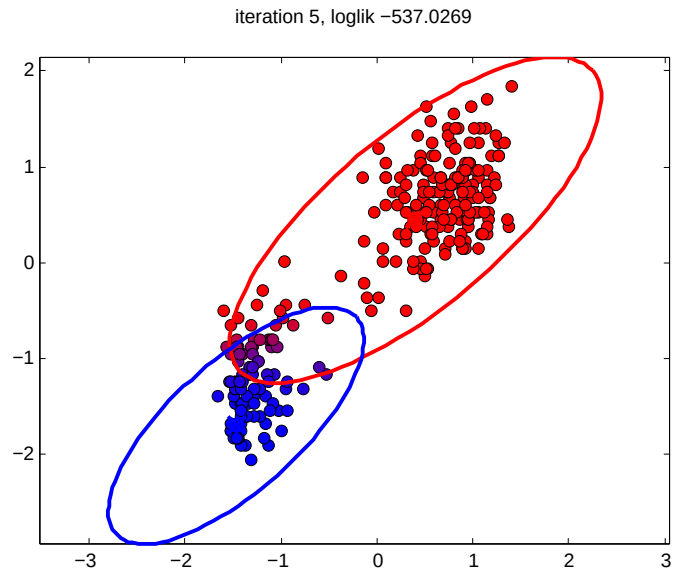


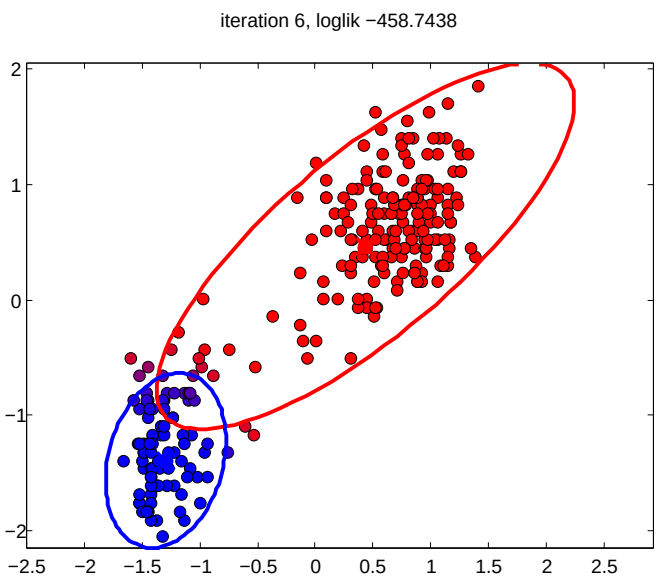


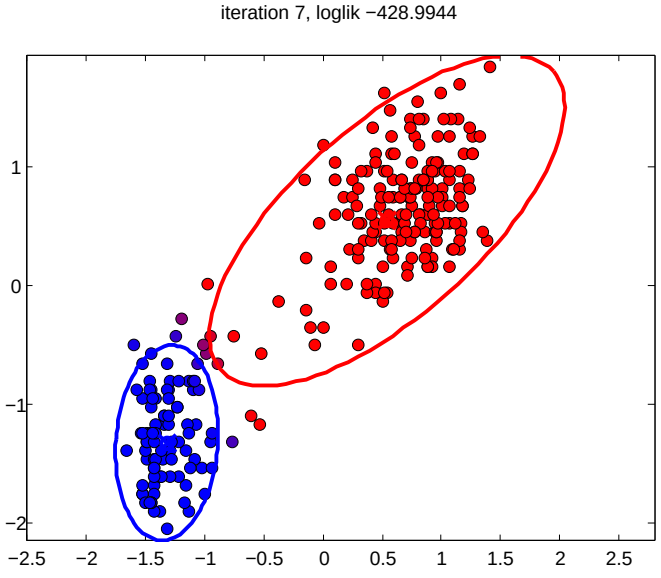


Example 2 of 4

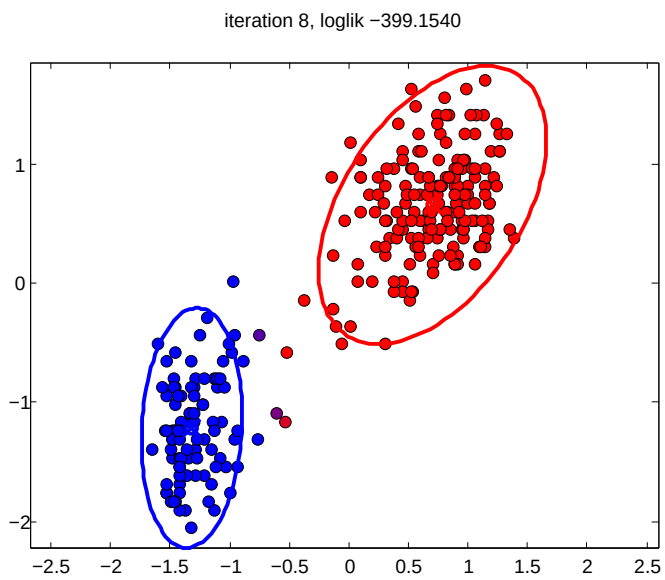


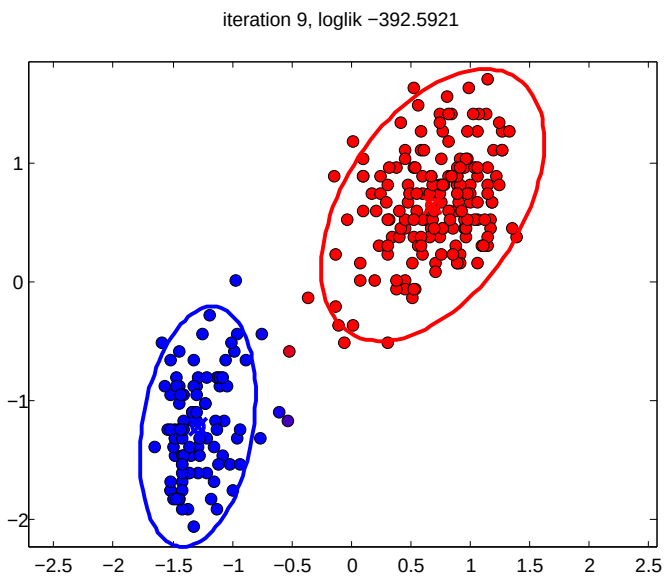


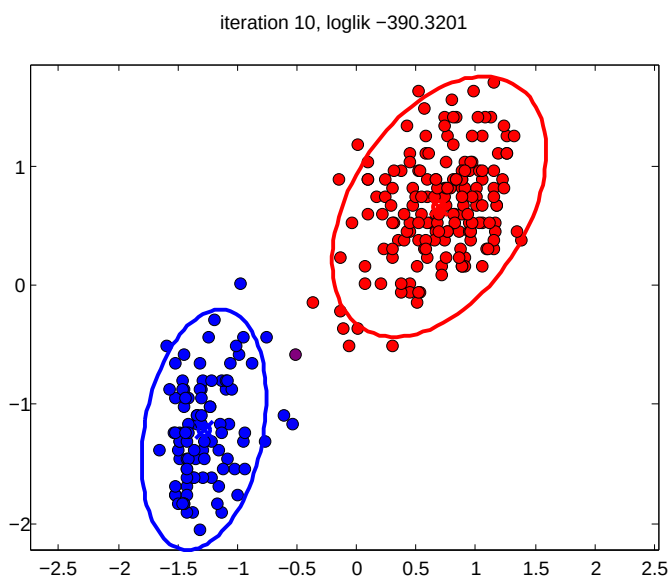


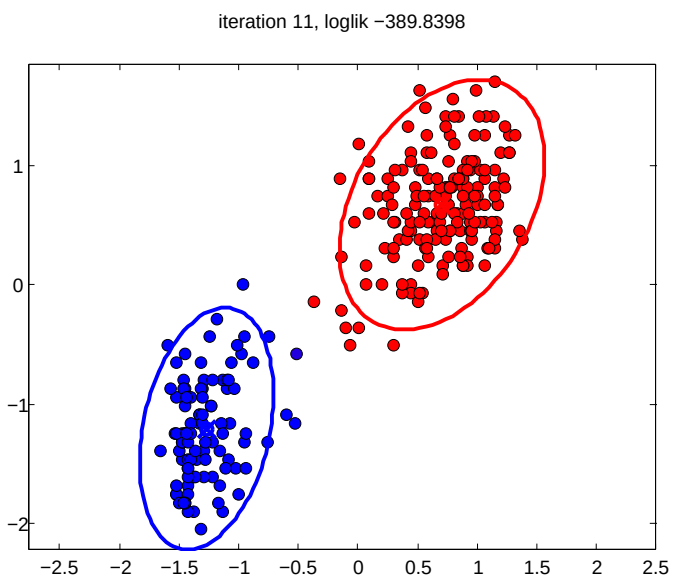


Example 3 of 4

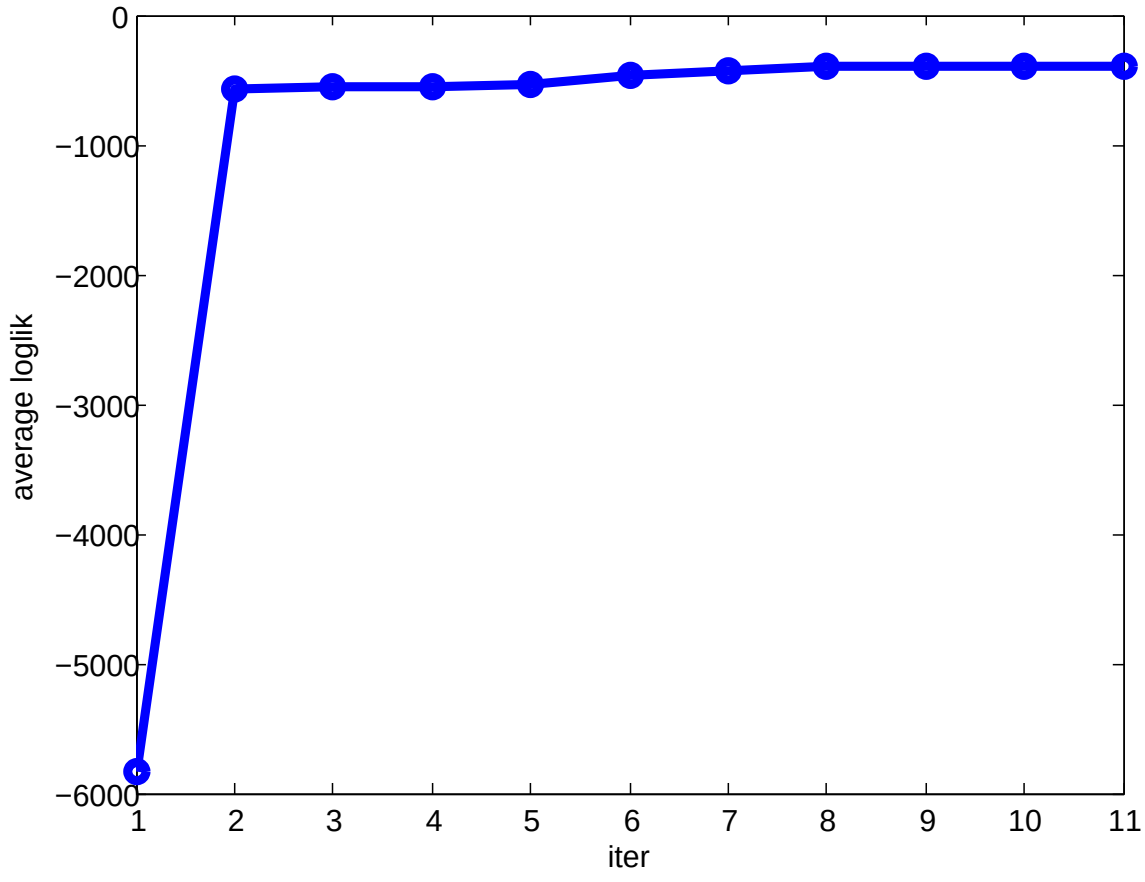








Example 4 of 4, Convergence



K -means algorithm

- K - means algorithm is guaranteed to converge.
- Clustering depend on the particular initialization. Different runs may produce different clusterings. Solution is not global.
- Centroids are the parameters of the model.
- K - means algorithm allows to discover latent structure of the data

K -means algorithm

- K - means algorithm is guaranteed to converge.
- Clustering depend on the particular initialization. Different runs may produce different clusterings. Solution is not global.
- Centroids are the parameters of the model.
- K - means algorithm allows to discover latent structure of the data.
- K - means algorithm works well when the data consists of well-separated Gaussians.
- K - means algorithm performs poorly on the data which does not resemble Gausssian at all.
- Number of classes K should be known or guessed.

K -means implementation in MATLAB environment

```
[idx,C,sumd,D] = kmeans(X,k,Name,Value)
```

- idx - returns cluster indexes for each point.
- C - returns centroids.
- sumd - for each cluster returns the sum of the distances from points to corresponding centroid.
- D - returns distance from each point to every centroid.
- X - initial data to cluster.
- k - number of clusters.
- Name refers to the name of the parameter name to be set.

  `'Distance'`
- Value is the value of the parameter to be set.

  `'cityblock'`

Gaussian

- One-dimensional
  - Do you remember a bell shaped curve?
  - Parameterized by mean μ and variance σ^2
  - Probability density function (pdf):

$$p(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-\mu)^2}{2\sigma^2}$$

- D-dimensional: Parameterized by mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ.

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \mid \Sigma \mid^{1/2} \exp\Big[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{\mu})\Big]$$

- Derive for the 2- and 3- dimensional cases.

Cluster Validation

- Internal Cluster Validation
  - Sum of square distances to centroids;
  - Intracluster to intercluster distance ratio;
  - Silhouette coefficient;
  - Probabilistic measure;
- External Cluster Validation, used when ground truth information is available.
  - Confusion matrix;
  - Cluster purity;
  - Gini index;

Intracluster to intercluster distance ratio

Sample r pairs of data points from the underlying data.

- Let P is the set of pairs that belong to the same cluster and Q is the set of remaining pairs.
- *Average intracluster distance:*

$$\text{Intra} = \frac{\sum_{\left(\bar{X}_i, \bar{X}_j\right) \in P} \text{dist}\left(\bar{X}_i, \bar{X}_j\right)}{|P|}$$

- *Average intercluster distance:*

$$\text{Inter} = \frac{\sum_{\left(\bar{X}_i, \bar{X}_j\right) \in Q} \text{dist}\left(\bar{X}_i, \bar{X}_j\right)}{|Q|}$$

- Smaller values of the ratio $\text{Intra}/\text{Inter}$ indicate better clustering behaviour.

Silhouette coefficient

- $\mathrm{Davg}_i^{\mathrm{in}}$ average distance of $\bar{X}_i$ to data points within its own cluster i.
- $\mathrm{Davg}_{i,j}^{\mathrm{out}}$ average distance of $\bar{X}_i$ to data points of cluster j.
- $\mathrm{Dmin}_i^{\mathrm{out}} = \min\{\mathrm{Davg}_{i,j}^{\mathrm{out}}\}$.
- Silhouette coefficient specific to the ith data point is defined as follows

$$S_i = \frac{\mathrm{Dmin}_i^{\mathrm{out}} - \mathrm{Davg}_i^{\mathrm{in}}}{\max\{\mathrm{Dmin}_i^{\mathrm{out}}, \mathrm{Davg}_i^{\mathrm{in}}\}}$$

- The overall silhouette coefficient is the average of data-point specific coefficients.
- The silhouette coefficient will take values from $(-1, 1)$. Large positive values indicate highly separated clusters.

Cluster Purity

- Let m_{ij} represent the number of data points from class (ground-truth cluster) i that are mapped to (algorithm determined) cluster j.
- Denote number of data points in true cluster i are by N_i,the number of data points in algorithm-determined cluster j by M_j.

$$N_i = \sum_{j=1}^{k_d} m_{ij}; \qquad M_j = \sum_{i=1}^{k_t} m_{ij};$$

- For a given algorithm-determined cluster j, the number of data points P_j in its dominant class is: $P_j = \max_i m_{ij}$.
- Purity index is defined

$$P_a = \frac{\sum_{j=1}^{k_d} P_j}{\sum_{j=1}^{k_d} M_j}.$$

Gini index

- Gini index for algorithm determined cluster j is defined:

$$G_j = 1 - \sum_{i=1}^{k_t} \left(\frac{m_{ij}}{M_j} \right)^2 .$$

- Average Gini index is defined as follows:

$$G = \frac{\sum_{j=1}^{k_d} G_j M_j}{\sum_{j=1}^{k_d} M_j}.$$

Relations to the Entropy

$$E_j = -\sum_{i=1}^{k_t} \left(\frac{m_{ij}}{M_j} \right) \log \left(\frac{m_{ij}}{M_j} \right).$$

$$E = \frac{\sum_{j=1}^{k_d} E_j M_j}{\sum_{j=1}^{k_d} M_j}.$$