Finite Mixture Models and Clustering

Mohamed Nadif

Université Paris Descartes, France

Outline

1. **Summary**
   - Clustering methods
2. **Introduction**
   - Mixture Approach
3. **Finite Mixture Model**
   - Definition of the model
   - Example
   - Different approaches
4. **ML approach**
   - EM algorithm
   - CEM algorithm
5. **Applications**
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. **Model Selection**
7. **Conclusion**

Advantages and Drawbacks of AHC and k-means

- k-means is faster than AHC
- Unlike k-means, in the merge process of AHC, once a cluster is formed, it does not undo what was previously, then no modification of clusters or permutations of objects are possible
- Unlike k-means, AHC does not require the knowledge of the number of clusters
- k-means depends on the initialization of the algorithm, The user must run it several times and choose the best result corresponding to the smallest value of W. The user can propose an initialization and in this case Run k-means one time

SOM

- A neuron k is characterized by the weight vector $\boldsymbol{\mu}_k$
- Description of the basic SOM
  - Choose the size of the grid initialization of the neurons: $\boldsymbol{\mu}_k^{(0)}$
  - Choose an object $\boldsymbol{x}_i^{(c+1)}$
  - Research of the winner k^*, $k^* = \text{argmin}_k \, ||\boldsymbol{x}_i^{(c+1)} - \boldsymbol{\mu}_k^{(c)}||$
  - The update of the weight vectors concern k^* and all neurons near of k^*

$$\boldsymbol{\mu}_k^{(c+1)} = \boldsymbol{\mu}_k^{(c)} + \varepsilon(t)h(k^*,\ell)(\boldsymbol{x}_i^{(c+1)} - \boldsymbol{\mu}_k^{(c)})$$

Outline

1. Summary
   - Clustering methods
2. Introduction
   - Mixture Approach
3. Finite Mixture Model
   - Definition of the model
   - Example
   - Different approaches
4. ML approach
   - EM algorithm
   - CEM algorithm
5. Applications
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. Model Selection
7. Conclusion

Classical clustering methods

- Clustering methods hierarchical and nonhierarchical methods have advantages and disadvantages
- Disadvantages. They are for the most part heuristic techniques derived from empirical methods
- Difficulties to take into account the characteristics of clusters (shapes, proportions, volume etc.)
- Geometrical approach: Clustering with "adaptives" distances: $d_{M_{\boldsymbol{k}}}(x, y) = ||x - y||_{M_{\boldsymbol{k}}}$
- In fact, the principal question "does it exist a model ?"

Mixture Approach

- MA have attracted much attention in recent years
- Is undoubtedly a very useful contribution to clustering
  1. It offers considerable flexibility
  2. provides solutions to the problem of the number of clusters
  3. Its associated estimators of posterior probabilities give rise to a fuzzy or hard clustering using the a MAP
  4. It permits to give a sense to certain classical criteria
- Finite Mixture Models by (McLachlan and Peel, 2000)

Outline

1. Summary
   - Clustering methods
2. Introduction
   - Mixture Approach
3. **Finite Mixture Model**
   - Definition of the model
   - Example
   - Different approaches
4. ML approach
   - EM algorithm
   - CEM algorithm
5. Applications
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. Model Selection
7. Conclusion

Definition of the model

- In model-based clustering it is assumed that the data are generated by a mixture of underlying probability distributions, where each component k of the mixture represents a cluster. Thus, the data matrix is assumed to be an i.i.d sample $\mathbf{x}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_n)$ where $\boldsymbol{x}_i=(x_{i1},\dots,x_{ip})\in\mathbb{R}^p$ from a probability distribution with density

$$f(\boldsymbol{x}_i;\boldsymbol{\theta})=\sum_k \pi_k\varphi_k(\boldsymbol{x}_i;\boldsymbol{\alpha}_k),$$

where

- $\varphi_k(.\,;\boldsymbol{\alpha}_k)$ is the density of an observation $\boldsymbol{x}_i$ from the k-th component
- $\boldsymbol{\alpha}_k$'s are the corresponding class parameters. These densities belong to the same parametric family
- The parameter π_k is the probability that an object belongs to the k-th component
- K, which is assumed to be known, is the number of components in the mixture

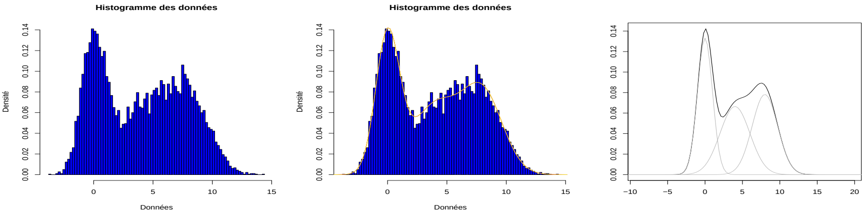
Gaussian mixture model in $\mathbb{R}^1$

- n=9000, d=1, K=3
- $\varphi(.,\boldsymbol{\alpha}_k)$ a Gaussian density $\boldsymbol{\alpha}_k = (m_k, s_k)$
- $\pi_1 = \pi_2 = \pi_3 = \frac{1}{3}$

The mixture density of the observed data **x** can be written as

$$f(\mathbf{x};\boldsymbol{\theta}) = \prod_i \sum_k \pi_k \prod_j \frac{1}{s_k\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x_i - m_k}{s_k})^2)$$

Mixture of 3 densities



Bernoulli mixture model

- The parameter of this model is the vector $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ containing the mixing proportions $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)$ and the vector $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, ..., \boldsymbol{\alpha}_K)$ of parameters of each component. The mixture density of the observed data $\mathbf{x}$ can be expressed as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \prod_i \sum_k \pi_k \varphi_k(\boldsymbol{x}_i; \boldsymbol{\alpha}_k).$$

- For instance, for binary data with $\boldsymbol{x}_i \in \{0,1\}^p$, using multivariate Bernoulli distributions for each component, the mixture density of the observed data $\mathbf{x}$ can be written as

$$f(\mathbf{x}; \boldsymbol{\theta}) = \prod_i \sum_k \pi_k \prod_j \alpha_{kj}^{x_{ij}} (1 - \alpha_{kj})^{1-x_{ij}}$$

where $x_{ij} \in \{0,1\}$, $\boldsymbol{\alpha}_k = (\alpha_{k1}, \ldots, \alpha_{kd})$ and $\alpha_{kj} \in (0,1)$

ML and CML approaches

- The problem of clustering can be studied in the mixture model using two different approaches: the maximum likelihood approach (ML) and the classification likelihood approach (CML)

  1. The ML approach (Day, 1699): It estimates the parameters of the mixture, and the partition on the objects is derived from these parameters using the maximum a posteriori principle (MAP). The maximum likelihood estimation of the parameters results in an optimization of the log-likelihood of the observed sample

  $$L_M(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}) = \sum_i \log \left(\sum_k \pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right)$$

  2. The CML approach (Symons, 1981): It estimates the parameters of the mixture and the partition *simultaneously* by optimizing the classification log-likelihood

  $$L_C(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \log f(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i,k} z_{ik} \log \left(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k) \right)$$

Outline

1. Summary
   - Clustering methods
2. Introduction
   - Mixture Approach
3. Finite Mixture Model
   - Definition of the model
   - Example
   - Different approaches
4. **ML approach**
   - EM algorithm
   - CEM algorithm
5. Applications
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. Model Selection
7. Conclusion

Introduction of EM

- Much effort has been devoted to the estimation of parameters for the mixture model
- Pearson used the method of moments to estimate $\boldsymbol{\theta} = (m_1, m_2, s_1^2, s_2^2, \pi)$ of a unidimensional Gaussian mixture model with two components

  $$f(\boldsymbol{x}_i; \boldsymbol{\theta}) = \pi\varphi(\boldsymbol{x}_i; m_1, s_1^2) + (1-\pi)\varphi(\boldsymbol{x}_i; m_2, s_2^2)$$

  required to solve polynomial equations of degree nine
- Generally, the appropriate method used in this context is the EM algorithm (Dempster et al., 1977). Two steps Estimation and Maximization
- This algorithm can be applied in different contexts where the model depends on unobserved latent variables. In mixture context $\mathbf{z}$ represents this variable. It denotes which $\boldsymbol{x}_i$ is from. Then we note $\mathbf{y} = (\mathbf{x}, \mathbf{z})$ the complete data.
- Starting from the relation between the densities

  $$f(\mathbf{y}, \boldsymbol{\theta}) = f((\mathbf{x}, \mathbf{z}); \boldsymbol{\theta})) = f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}) f(\mathbf{x}; \boldsymbol{\theta})$$

  we have

  $$\log(f(\mathbf{x}; \boldsymbol{\theta})) = \log(f(\mathbf{y}, \boldsymbol{\theta})) - \log(f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}))$$

  or

  $$L_M(\boldsymbol{\theta}) = L_C(\mathbf{z}; \boldsymbol{\theta}) - \log f(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta})$$

Principle of EM

- Objective: Maximization of $L_M(\boldsymbol{\theta})$
- EM rets on the hypothesis that maximizing L_C is simple
- An iterative procedure based on the conditional expectation of $L_M(\boldsymbol{\theta})$ for a value of the current parameter $\boldsymbol{\theta}'$

$$L_M(\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}') - H(\boldsymbol{\theta}|\boldsymbol{\theta}')$$

where $Q(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathbb{E}(L_C(\boldsymbol{\theta}|\mathbf{x},\boldsymbol{\theta}'))$ and $H(\boldsymbol{\theta}|\boldsymbol{\theta}') = \mathbb{E}(\log f(\mathbf{y}|\mathbf{x};\boldsymbol{\theta})|\mathbf{x},\boldsymbol{\theta}')$

- Using the Jensen inequality (Dempster et al;, 1977) for fixed $\boldsymbol{\theta}'$ we have $\forall\boldsymbol{\theta}, H(\boldsymbol{\theta}|\boldsymbol{\theta}') \leq H(\boldsymbol{\theta}'|\boldsymbol{\theta}')$ This inequality can proved also

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}') - H(\boldsymbol{\theta}'|\boldsymbol{\theta}') = \sum_{z\in\mathcal{Z}} f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}') \log \frac{f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')}$$

As $\log(x) \leq x - 1$, we have

$$\log \frac{f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')} \leq \frac{f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})}{f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}')} - 1$$

then

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}') - H(\boldsymbol{\theta}'|\boldsymbol{\theta}') \leq \sum_{z\in\mathcal{Z}} f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}) - \sum_{z\in\mathcal{Z}} f(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}') = 1 - 1 = 0$$

$Q(\theta|\theta')$

- The value θ maximizing maximization $Q(\theta|\theta')$ satisfies the relation $Q(\theta|\theta') \geq Q(\theta'|\theta')$ and,

$$L_M(\theta) = Q(\theta|\theta') - H(\theta|\theta') \geq Q(\theta'|\theta') - H(\theta'|\theta') = L_M(\theta')$$

- In mixture context

$$Q(\theta|\theta') = \mathbb{E}(L_C(\theta|\mathbf{x},\theta')) = \sum_{i,k} \mathbb{E}(z_{ik}|\mathbf{x},\theta') \log(\pi_k f(\boldsymbol{x}_i;\alpha_k))$$

- Note that $\mathbb{E}(z_{ik}|\mathbf{x},\theta') = p(z_{ik} = 1|\mathbf{x},\theta')$
As the conditional distribution of the missing data $\mathbf{z}$ given the observed values :

$$f(\mathbf{z}|\mathbf{x};\theta) = \frac{f(\mathbf{x},\mathbf{z};\theta)}{f(\mathbf{x};\theta)} = \frac{f(\mathbf{x}|\mathbf{z};\theta)f(\mathbf{z};\theta)}{f(\mathbf{x};\theta)}$$

we have

$$p(z_{ik} = 1|\mathbf{x},\theta') = s_{ik} = \frac{\pi_k\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)}{f(\mathbf{x};\theta)} = \frac{\pi_k\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_k)}{\sum_\ell \pi_\ell\varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_\ell)}$$

The steps of EM

- The EM algorithm involves constructing, from an initial $\boldsymbol{\theta}^{(0)}$, the sequence $\boldsymbol{\theta}^{(c)}$ satisfying
$$\boldsymbol{\theta}^{(c+1)} = \operatorname{argmax} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})$$
and this sequence causes the criterion $L_M(\boldsymbol{\theta})$ to grow The EM algorithm takes the following form
  - Initialize by selecting an initial solution $\boldsymbol{\theta}^{(0)}$
  - Repeat the two steps until convergence
    1. E-step: compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(c)})$. Note that in the mixture case this step reduces to the computation of the conditional probabilities $s_{\boldsymbol{ik}}^{(\boldsymbol{c})}$
    2. M-step: compute $\boldsymbol{\theta}^{(\boldsymbol{c}+\boldsymbol{1})}$ maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(\boldsymbol{c})})$. This leads to $\pi_{\boldsymbol{k}}^{(\boldsymbol{c}+\boldsymbol{1})} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i}} s_{\boldsymbol{ik}}^{(\boldsymbol{c}+\boldsymbol{1})}$ and the exact formula for the $\alpha_{\boldsymbol{k}}^{(\boldsymbol{c}+\boldsymbol{1})}$ will depend on the involved parametric family of distribution probabilities

Properties of EM

- Under certain conditions, it has been established that EM always converges to a local likelihood maximum
- Simple to implement and it has good behavior in clustering and estimation contexts
- Slow in some situations

An other interpretation of EM

Hathaway interpretation of EM : classical mixture model context

- EM = alternated maximization of the fuzzy clustering criterion

$$F_C(\mathbf{s},\boldsymbol{\theta}) = L_C(\mathbf{s};\boldsymbol{\theta}) + H(\mathbf{s})$$

  - $\mathbf{s} = (s_{ik})$: fuzzy partition
  - $L_C(\mathbf{s},\boldsymbol{\theta}) = \sum_{i,k} s_{ik} \log(\pi_k \varphi(\boldsymbol{x}_i;\boldsymbol{\alpha}_k))$: fuzzy classification log-likelihood
  - $H(\mathbf{s}) = -\sum_{i,k} s_{ik} \log s_{ik}$: entropy function

Algorithm

- Maximizing F_C w.r. to $\mathbf{s}$ yields the E step
- Maximizing F_C w.r. to $\boldsymbol{\theta}$ yields the M step

Stochastic EM "SEM", (Celeux and Diebolt, 1985)

Steps of SEM

- S-step between E-step and M-step
- In CEM (C-step), In SEM (S-step)
  - E-step compute the posterior probabilities
  - S-step This stochastic step consists to look for the partition $\bar{\mathbf{z}}$. Each object i is assigned to the kth component. the parameter k is selected according to the multinomial distribution $(s_{i1}, \ldots, s_{iK})$
  - M-step As the CEM algorithm this step is based on $\bar{\mathbf{z}}$

Advantages and Disadvantages of SEM

- It gives good results when the size of data is large enough
- It can be used even if the number of clusters is unknown. It suffices to fix K to K_{max} the maximum number of clusters and this number can be reduced when the a cluster has a number of objects so lower that the estimation of parameters is not possible. For example when the cardinality of a cluster is less than a threshold, we run SEM with $(K-1)$
- It can avoid the problem of initialization and other problems of EM
- Instability of the results. Solution: SEM (for estimation of paremetrs et the number of clusters), The obtained results are used by EM

Stochastic Annealing EM "SAEM" (Celeux and Diebolt, 1992)

Steps of SEM

- The aim of the SAEM is to reduce the "part" of random in estimations of the parameters
- SAEM is based on SEM and EM
- Solution
  - E-step: Idem for EM, SEM
  - S-step: Idem for SEM
  - M-step: The compute of parameters depends on this expression:

$$\boldsymbol{\theta}^{(t+1)} = \gamma^{(t+1)}\boldsymbol{\theta}_{SEM}^{(t+1)} + (1-\gamma^{(t+1)})\boldsymbol{\theta}_{EM}^{(t+1)}$$

The initial value of $\gamma = 1$ and decreases until 0.

CEM algorithm

- In the CML approach the partition is added to the parameters to be estimated. The maximum likelihood estimation of these new parameters results in an optimization of the complete data log-likelihood. This optimization can be performed using the following Classification EM (CEM) algorithm (Celeux and Govaert, 1992), a variant of EM, which converts the s_{ik}'s to a discrete classification in a C-step before performing the M-step:
  - E-step: compute the posterior probabilities $s_{ik}^{(c)}$.
  - C-step: the partition $\mathbf{z}^{(c+1)}$ is defined by assigning each observation $\boldsymbol{x}_i$ to the cluster which provides the maximum current posterior probability.
  - M-step: compute the maximum likelihood estimate $(\pi_k^{(c+1)}, \boldsymbol{\alpha}_k^{(c+1)})$ using the k-th cluster. This leads to $\pi_k^{(c+1)} = \frac{1}{n}\sum_i z_{ik}^{(c+1)}$ and the exact formula for the $\boldsymbol{\alpha}_k^{(c+1)}$ will depend on the involved parametric family of distribution probabilities

Properties of CEM

- Simple to implement and it has good practical behavior in clustering context
- Faster than EM and scalable
- Some difficulties when the clusters are not well separated

Link between CEM and the dynamical clustering methods

Dynamical clustering method	The CEM algorithm
Assignation-step $z_k = \{i; d(x_i, a_k) \leq d(x_i, a'_k); k' \neq k\}$	E-step Compute $s_{ik} \propto \pi_k \varphi(x_i, \alpha_k)$ C-step $z_k = \{i; s_{ik} \geq s_{ik'}; k' \neq k\}$ $z_k = \{i; -log(\pi_k \varphi(x_i, \alpha_k)) \leq -log(\pi_k \varphi(x_i, \alpha'_k)); k' \neq k\}$
Representation-step Compute the center a_k of each cluster	M-step Compute the π_k's and α_k

Density and distance

- When the proportions are supposed equal we can propose a *distance D* by

$$D(\mathbf{x}_i, \mathbf{a}_k) = -log(\varphi(\mathbf{x}_i, \boldsymbol{\alpha}_k))$$

Outline

1. Summary
   - Clustering methods
2. Introduction
   - Mixture Approach
3. Finite Mixture Model
   - Definition of the model
   - Example
   - Different approaches
4. ML approach
   - EM algorithm
   - CEM algorithm
5. Applications
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. Model Selection
7. Conclusion

The Gaussian model

- The density can be written as: $f(\boldsymbol{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k \varphi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\Sigma_k)$ where

$$\varphi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\Sigma_k) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} exp\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\boldsymbol{x}_i-\mu_k)\}$$

- Spectral decomposition of the variance matrix

$$\Sigma_k = \lambda_k D_k A_k D_k^T$$

  - $\lambda_k = |\Sigma_k|^{1/p}$ positive real represents the volume of the kth component
  - $A_k = Diag(a_{k1},\ldots,a_{kp})$ formed by the normalized eigenvalues in decreasing order $|A_k| = 1$. It defines the shape of the kth cluster
  - D_k formed by the eigenvectors. It defines the direction of the kth cluster

- Example in $\mathbb{R}^2$, D_k is a rotation, and A_k is diagonal matrix, the equidensity ellipse of the distribution depends on the center $\boldsymbol{\mu}_k$, semimajor axis and semiminor axis $\sqrt{\lambda_k a}$ and $\sqrt{\lambda_k/a}$

$$D_k = \begin{pmatrix} cos(\alpha) & sin(\alpha) \\ -sin(\alpha) & cos(\alpha) \end{pmatrix} A_k = \begin{pmatrix} a & 0 \\ 0 & 1/a \end{pmatrix}$$

Different Gaussian models

- The Gaussian mixture depends on: proportions, centers, volumes, shapes and Directions then different models can be proposed
- In the following models proportions can be assumed equal or not

1. Spherical models: $A_k = I$ then $\Sigma_k = \lambda_k I$. Two models $[\lambda I]$ and $[\lambda_k I]$
2. Diagonal models: no constraint on A_k but D_k is a permutation matrix with $B_k = D_k A_k D_k^T$ such as $|B_k| = 1$, Σ_k is diagonal. Four models $[\lambda B]$, $[\lambda_k B]$, $[\lambda B_k]$ and $[\lambda_k B_k]$
3. General models: the eight models assuming equal or not volumes, shapes and directions $[\lambda DAD_k^T]$, $[\lambda DAD^T]$, $[\lambda DA_k D^T]$, $[\lambda_k DA_k D^T]$, $[\lambda D_k AD_k^T]$,$[\lambda_k D_k AD_k^T]$, $[\lambda D_k A_k D_k^T]$ and $[\lambda_k D_k A_k D_k^T]$

- Finally we have 28 models, we will study the problem of the choice of the models

CEM

- In clustering step, each $\boldsymbol{x}_i$ is assigned to the cluster maximizing $s_{ik} \propto \pi_k\varphi(\boldsymbol{x}_i; \mu_k, \Sigma_k)$ or equivalently the cluster that minimizes

$$-\log(\pi_k\varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)) = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log(\pi_k) + cste$$

- From density to Distance (or dissimilarity), $\boldsymbol{x}_i$ is assigned to the cluster according the following dissimilarity

$$d_{\Sigma_{\boldsymbol{k}}^{-1}}(\boldsymbol{x}_i; \boldsymbol{\mu}_k) + \log|\Sigma_k| - 2\log(\pi_k)$$

where $d_{\Sigma_{\boldsymbol{k}}^{-1}}(\boldsymbol{x}_i; \mu_k) = (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T\Sigma_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)$ is the Mahanalobis distance

- Note that when the proportions are supposed equal and the variances identical, the assignation is based only on

$$d^2_{\Sigma_{\boldsymbol{k}}^{-1}}(\boldsymbol{x}_i; \boldsymbol{\mu}_k)$$

- When the proportions are supposed equal and for the spherical model $[\lambda I]$ $(\Sigma_k = I)$, one uses the usual euclidean distance

$$d^2(\boldsymbol{x}_i; \boldsymbol{\mu}_k)$$

Description of CEM

- E-step: classical, C-step: Each cluster z_k i formed by using $d^2(\boldsymbol{x}_i; \boldsymbol{\mu}_k)$
- M-step: Given the partition $\mathbf{z}$, we have to determine the parameter $\boldsymbol{\theta}$ maximizing

$$L_C(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i,k} z_{ik} \log\left(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)\right) = \sum_k \sum_{i \in z_{\boldsymbol{k}}} \log\left(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)\right)$$

For the Gaussian model

$$-\frac{1}{2} \sum_k \left(\sum_{i \in z_{\boldsymbol{k}}} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) + \# z_k \log|\Sigma_k| - 2\# z_k \log(\pi_k) \right)$$

- The parameter $\boldsymbol{\mu}_k$ is thus necessary the center $\boldsymbol{\mu}_k = \frac{\sum_{\boldsymbol{i} \in \boldsymbol{z_k}} \boldsymbol{x_i}}{\# z_{\boldsymbol{k}}}$
- The proportions satisfy $\pi_k = \frac{\# z_{\boldsymbol{k}}}{n}$
- The parameters must then for the general model

$$F(\Sigma_1, \ldots, \Sigma_K) = \sum_k (\text{trace}(W_k \Sigma_k^{-1}) + \# z_k \log|\Sigma_k|)$$

where $W_k = \sum_{i \in z_{\boldsymbol{k}}} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$

Consequence for the spherical model $[\lambda I]$

- Exercise: The function to maximize for the model $[\lambda I]$ becomes

$$F(\lambda) = \frac{1}{\lambda}\text{trace}(W) + np\log(\lambda)$$

where $W = \sum_k W_k$

With $\lambda = \frac{\text{trace}(W)}{np}$ maximizing $F(\lambda)$, the classification log-likelihood becomes

$$L_C(\boldsymbol{\theta}) = -\frac{np}{2}\text{trace}(W) + cste = -\frac{np}{2}W(\mathbf{z}) + cste$$

- Maximizing L_C is equivalent to minimize the SSQ criterion minimized by the $k$means algorithm
- Interpretation
  - The use of the model $[\lambda I]$ assumes that the clusters are spherical having the same proportion and the same volume
  - The CEM is therefore an extension of the $k$means

Description of EM

- E-step: classical
- M-step: we have to determine the parameter $\boldsymbol{\theta}$ maximizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}')$ taking the following form

$$L_C(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = \sum_{i,k} s_{ik} \log\left(\pi_k \varphi(\boldsymbol{x}_i; \boldsymbol{\alpha}_k)\right)$$

For the Gaussian model

$$-\frac{1}{2} \sum_{i,k} \left(s_{ik}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) + s_{ik} \log|\Sigma_k| - 2 s_{ik} \log(\pi_k) \right)$$

- The parameter $\boldsymbol{\mu}_k$ is thus necessary the center $\boldsymbol{\mu}_k = \frac{\sum_i s_{ik} x_i}{\sum_i s_{ik}}$
- The proportions satisfy $\pi_k = \frac{\sum_i s_{ik}}{n}$
- The parameters Σ_k must then minimize

$$F(\Sigma_1, \ldots, \Sigma_K) = \sum_k (\text{trace}(W_k \Sigma_k^{-1}) + \# z_k \log|\Sigma_k|)$$

where $W_k = \sum_{i \in z_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k)$

Binary data

- For binary data, considering the conditional independence model (independence for each component), the mixture density of the observed data $\mathbf{x}$ can be written as

$$f(\mathbf{x};\boldsymbol{\theta}) = \prod_i \sum_k \pi_k \prod_j \alpha_{kj}^{x_{ij}} (1-\alpha_{kj})^{1-x_{ij}}$$

  where $x_{ij} \in \{0,1\}$, $\boldsymbol{\alpha}_k = (\alpha_{k1},\ldots,\alpha_{kp})$ and $\alpha_{kj} \in (0,1)$
- Latent Class Model
- The different steps of EM algorithm

  1. E-step: compute s_{ik}
  2. M-step: $\alpha_k^j = \frac{\sum_i s_{ik} x_i^j}{\sum_i s_{ik}}$ and $\pi_k = \frac{\sum_i s_{ik}}{n}$

- The different steps of CEM algorithm

  1. E-step: compute s_{ik}
  2. C-step: compute $\mathbf{z}$
  3. M-step: $\alpha_k^j = \frac{\sum_i z_{ik} x_i^j}{\sum_i z_{ik}} = \%1$ and $\pi_k = \frac{\# z_k}{n}$

Parsimonious model

- As for the Gaussian, several parsimonious models can be proposed

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k \prod_j \varepsilon_{kj}^{|x_{\mathbf{ij}}-a_{\mathbf{kj}}|}(1-\varepsilon_{kj})^{1-|x_{\mathbf{ij}}-a_{\mathbf{kj}}|}$$

where

$$\begin{cases} a_{kj} = 0, \varepsilon_{kj} = a_{kj} & \text{if } \alpha_{kj} < 0.5 \\ a_{kj} = 1, \varepsilon_{kj} = 1 - a_{kj} & \text{if } \alpha_{kj} > 0.5 \end{cases}$$

- The parameter $\boldsymbol{\alpha}_k$ is replaced by the two parameters a_k and ε_k
  - The binary vector a_k represents the center of the cluster z_k, each a_{kj} indicates the most frequent binary value
  - The binary vector $\varepsilon_k \in]0, 1/2[^p$ represents the degrees of heterogeneity of the cluster z_k, each ε_{kj} represents the probability of j to have the value different from that of the center,
    - $p(x_{\mathbf{ij}} = 1|a_{\mathbf{kj}} = 0) = p(x_{\mathbf{ij}} = 0|a_{\mathbf{kj}} = 1) = \varepsilon_{\mathbf{kj}}$
    - $p(x_{\mathbf{ij}} = 0|a_{\mathbf{kj}} = 0) = p(x_{\mathbf{ij}} = 1|a_{\mathbf{kj}} = 1) = 1 - \varepsilon_{\mathbf{kj}}$
- 8 Models assuming proportions equal or not : $[\varepsilon_{kj}]$, $[\varepsilon_k]$, ε_j, $[\varepsilon]$

Binary data matrix and reorganized data matrix

	a	b	c	d	e
1	1	0	1	0	1
2	0	1	0	1	0
3	1	0	0	0	0
4	1	0	1	0	0
5	0	1	0	1	1
6	0	1	0	0	1
7	0	1	0	0	0
8	1	0	1	0	1
9	1	0	0	1	0
10	0	1	0	1	0

	a	b	c	d	e
1	1	0	1	0	1
4	1	0	1	0	0
8	1	0	1	0	1
2	0	1	0	1	0
5	0	1	0	1	1
6	0	1	0	0	1
10	0	1	0	1	0
3	1	0	0	0	0
7	0	1	0	0	0
9	1	0	0	1	0

Centers $\mathbf{a}_k$ and Degree of heterogeneity ε_k

	a	b	c	d	e
$\mathbf{a}_1$	1	0	1	0	1
$\mathbf{a}_2$	0	1	0	1	0
$\mathbf{a}_3$	1	0	0	0	0

	a	b	c	d	e
ε_1	0	0	100	0	0.33
ε_2	0	0	100	0.25	0.5
ε_3	0.33	0.33	0	0.33	0

CEM for the simplest model $[\varepsilon]$

- Exercise: When the proportions are supposed equal The classification log-likelihood to maximize

$$L_C(\theta) = L(\theta; \mathbf{x}, \mathbf{z}) = \log(\frac{\varepsilon}{1-\varepsilon}) \sum_k \sum_{i \in z_\mathbf{k}} d(\boldsymbol{x}_i, \mathbf{a}_k) + np \log(1-\varepsilon)$$

  where $d(\boldsymbol{x}_i, \mathbf{a}_k) = \sum_j |x_{ij} - a_{kj}|$

- The parameter ε is fixed for each cluster and for each variable, as $(\log(\frac{\varepsilon}{1-\varepsilon}) \leq 0)$ this maximization leads to the minimization of

$$W(\mathbf{z}, \mathbf{a}) = \sum_k \sum_{i \in z_\mathbf{k}} d(\boldsymbol{x}_i, \mathbf{a}_k)$$

- Exercise: The CEM algorithm is equivalent to the dynamical clustering method

CEM and EM for the other models

- Exercise: Describe the different steps of CEM for the models $[\varepsilon_j]$, $[\varepsilon_k]$ and $[\varepsilon_{kj}]$
- Exercise: Deduce the different steps of EM for these models

Nominal categorical data

- Categorical data are a generalization of binary data
- Generally this kind of data are represented by a *complete disjunctive table* where the categories are represented by their indicators
- A variable j with h categories is represented by a binary vector such as

$$\begin{cases} x_i^{jh} = 1 & \text{if } i \text{ takes the categorie } h \text{ for} j \\ x_i^{jh} = 0 & \text{otherwise} \end{cases}$$

- The probability of the mixture can be written

$$f(\mathbf{x}_i;\boldsymbol{\theta}) = \sum_k \pi_k \prod_{j,h} (\alpha_k^{jh})^{x_{\mathbf{ij}}}$$

where α_k^{jh} is the probability that the variable j takes the categorie h when an object belongs to the cluster k.

Notation

- $d_k^{jh} = \sum_{i \in z_k} x_i^{jh}$
- $d^{jh} = \sum_i x_i^{jh}$
- $d_k = \sum_{j,h} d_k^{jh}$
- $d = \sum_k d_k = \sum_{k,j,h} x_i^{jh} = np$

Example

	a	b
1	1	2
2	3	2
3	2	3
4	1	1
5	1	2
6	3	2
7	3	3
8	1	1
9	2	2
10	2	3

	a1	a2	a3	b1	b2	b3
1	1	0	0	0	1	0
2	0	0	1	0	1	0
3	0	1	0	0	0	1
4	1	0	0	1	0	0
5	1	0	0	0	1	0
6	0	0	1	0	1	0
7	0	0	1	0	0	1
8	1	0	0	1	0	0
9	0	1	0	0	1	0
10	0	1	0	0	0	1

	a1	a2	a3	b1	b2	b3
3	0	1	0	0	0	1
7	0	0	1	0	0	1
9	0	1	0	0	1	0
10	0	1	0	0	0	1
1	1	0	0	0	1	0
4	1	0	0	1	0	0
5	1	0	0	0	1	0
8	1	0	0	1	0	0
2	0	0	1	0	1	0
6	0	0	1	0	1	0

- $d_1^{a1} = 0, d_1^{a2} = 3,\ d_1^{a3} = 1,\ d_1^{b1} = 0, d_1^{b2} = 1,\ d_1^{b3} = 3$
- $d_1 = 8,\ d_2 = 8,\ d_3 = 4$
- $d = 8 + 8 + 4 = 10 \times 2$

Interpretation of the model

- The different steps of EM algorithm
  1. E-step: compute s_{ik}
  2. M-step: $\alpha_k^{jh} = \frac{\sum_i s_{ik} x_i^{jh}}{\sum_i s_{ik}}$ and $\pi_k = \frac{\sum_{i,k} s_{ik}}{n}$
- The different steps of CEM algorithm
  1. E-step: compute s_{ik}
  2. C-step: compute $\mathbf{z}$
  3. M-step (Exercise) : $\alpha_k^{jh} = \frac{\sum_i z_{ik} x_i^{jh}}{\sum_i z_{ik}} = \frac{d_k^{jh}}{\# z_k}$ and $\pi_k = \frac{\# z_k}{n}$

Interpretation of the model

- The classification log-likelihood can be written as

$$L_C(\boldsymbol{\theta}) = \sum_{k,j,h} d_k^{jh} \log(\alpha_k^{jh}) + \sum_k \# z_k \log(\pi_k)$$

- When the proportions are supposed equal, the restricted likelihood

$$L_{CR}(\boldsymbol{\theta}) = \sum_{k,j,h} d_k^{jh} \log(\alpha_k^{jh})$$

- Given $\alpha_k^{jh} = \frac{d_{\mathbf{k}}^{\mathbf{jh}}}{\# z_{\mathbf{k}}}$, it can be shown that the CEM algorithm maximizes $H(\mathbf{z})$

$$H(\mathbf{z}) = \sum_{k,j,h} \frac{d_k^{jh}}{d} \log \frac{d_k^{jh} d}{d_k d^{jh}}$$

  This expression is very close to

$$\chi^2(\mathbf{z}) = \sum_{k,j,h} \frac{(d_k^{jh} d - d_k d^{jh})^2}{d_k d^{jh} d}$$

- To assume that the date derive form the latent class model where the proportions are assumed equal is approximatively equivalent to use the χ^2 criterion

Parsimonious model

- Number of the parameters in latent class model is equal $(K-1)+K*\sum_j m_j - 1$ where m_j is the number of categories of j
- This number is smaller than $\prod_j m_j$ required by the complete log-linear model, example ($p=10$, K=5, $m_j=4$ for each j), this number is equal to $(5-1)+5*(40-10)=154$
- This number can reduced by using parsimonious model by imposing constraints on the paremetre α_{kj}. Instead to have a probability for each categorie, we associate for a categorie of j having the same of value that the center for j the probability $(1-\varepsilon_{kj})$ and the others categories the probability $\varepsilon_{kj}/(m_j-1)$
- Then the distribution depends on $\mathbf{a}_k$ and ε_k defined by

$$\begin{cases} (1-\varepsilon_{kj}) & \text{for } x_i^j = a_k^j \\ \varepsilon_{kj}/(m_j-1) & \text{for } x_i^j \neq a_k^j \end{cases}$$

- The parametrization concerns only the variables instead of all categories, the number of parameters becomes $(K-1)+2Kp$
- This model is an extension of the Bernoulli model

The simplest model

- We assume that $(1-\varepsilon_{kj})$ does not depend the cluster k and the variable j

$$\begin{cases} (1-\varepsilon) & \text{for } x_i^j = a_k^j \\ \varepsilon/(m_j-1) & \text{for } x_i^j \neq a_k^j \end{cases}$$

- Exercise: The restricted classification log-likelihood takes the following form

$$L_{CR}(\boldsymbol{\theta}) = L(\boldsymbol{\theta};\mathbf{x},\mathbf{z}) = \sum_k \sum_{i\in z_{\mathbf{k}}} \left(\sum_j \log\left(\frac{\varepsilon}{1-\varepsilon}(m_j-1)\right)\delta(\boldsymbol{x}_i,\mathbf{a}_k) \right) + np\log(1-\varepsilon)$$

or,

$$L_{CR}(\boldsymbol{\theta}) = \sum_k \sum_{i\in z_{\mathbf{k}}} d(\boldsymbol{x}_i,\mathbf{a}_k) + np\log(1-\varepsilon)$$

where $d(\boldsymbol{x}_i,\mathbf{a}_k) = \sum_j \log(\frac{1-\varepsilon}{\varepsilon}(m_j-1))\delta(x_{ij},a_{kj})$

- If all variables have the same number of categories, the criterion to minimize is $\sum_k \sum_{i\in z_{\mathbf{k}}} d(\boldsymbol{x}_i,\mathbf{a}_k)$, why ?
- The CEM is an extension of k-modes

Contingency table

- As for categorical, we can associate a multinomial model
- See (Govaert and Nadif 2007)

Outline

1. Summary
   - Clustering methods
2. Introduction
   - Mixture Approach
3. Finite Mixture Model
   - Definition of the model
   - Example
   - Different approaches
4. ML approach
   - EM algorithm
   - CEM algorithm
5. Applications
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. **Model Selection**
7. Conclusion

Different approaches

- In Finite mixture model, the problem of the choice of the model include the problem of the number of clusters
- To simplify the problem, we distinguish the two problems and we consider the model fixed and K is unknown. Let be tow models M_A and M_B. $\mathbf{\Theta}(M_A)$ and $\mathbf{\Theta}(M_B)$ indicates the "domain" of free parameters. if $L_{max}(M) = L(\hat{\theta}_M)$ where $\hat{\theta}_M = \operatorname{argmax} L(\theta)$ then we have

  $$\mathbf{\Theta}(M_A) \subset \mathbf{\Theta}(M_B) \Rightarrow L_{max}(M_A) \leq L_{max}(M_B)$$

  For example $L_{max}[\pi_k \lambda_k I]_{K=2} \leq L_{max}[\pi_k \lambda_k I]_{K=3}$. Generally the likelihood increases with the number of clusters.
- First solution: Plot (Likelihood*number of clusters) and use the elbows
- Second solution: Minimize the classical criteria (Criteria in competition) taking this form

  $$C(M) = -2L_{max}(M) + \tau_C n_p(M)$$

  where n_p indicates the number of parameters of the model M, it represents the complexity of the model
- Different variants of this criterion AIC with $\tau_{AIC} = 2$, AIC3 with $\tau_{AIC} = 3$ and the famous

  $$BIC(M) = -2L_{max}(M) + \log(n) n_p(M)$$

Outline

1. Summary
   - Clustering methods
2. Introduction
   - Mixture Approach
3. Finite Mixture Model
   - Definition of the model
   - Example
   - Different approaches
4. ML approach
   - EM algorithm
   - CEM algorithm
5. Applications
   - Gaussian mixture model
   - Bernoulli mixture
   - Multinomial Mixture
6. Model Selection
7. **Conclusion**

Conclusion

- Finite mixture approach is interesting
- The CML approach gives interesting criteria and generalize the classical criteria
- The different variants of EM offer good solutions
- The choice of the model is performed by using the maximum likelihood penalized by the number of parameters