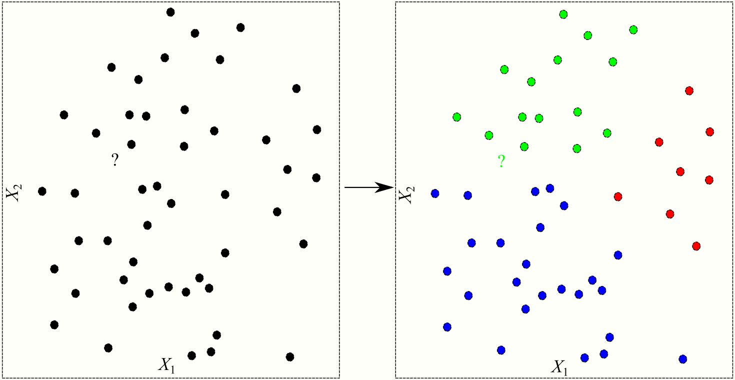
Fundamentals of Probabilistic Data Mining

Chapter III - Model-based clustering

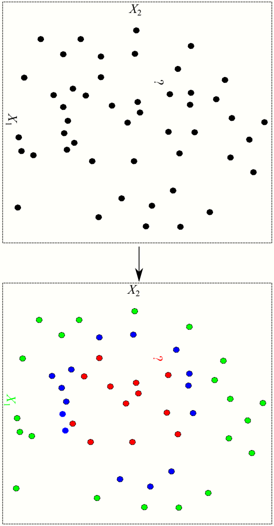


Clustering with 3 clusters

Mixture models

Clustering

- Data: points $(x_j)_{j=1,\ldots,n}$ in $\mathbb{R}^d$.
- Aim: find (maybe predict?) K clusters (K fixed here).
- Distance-based approaches: close points tend to be in the same cluster. No explicit assumption required. Clusters cannot be nested.
- Model-based approaches: let z_j be the cluster of x_j. If $z_i = z_j = k$ then x_i and z_j should have the same (conditional) distribution p_k.



Z is an unknown / latent variable, useful for clustering. (e.g. $z = 0 \to$ blue)

Mixture models

- Added value: to incorporate probabilistic constraints (clusters with different means, or same means but different covariances...)
- Example (latter case): $p(x|z=k) = \mathcal{N}(0, \sigma_k I)$, or $p(x|z=k) = \mathcal{N}(0, \sigma_k \Sigma)$, etc.

Definition 1 (McLachalan & Peel, 2000).

Let $\{p_\theta\}_{\theta \in \Theta}$ be a parametric family of distributions. $x \to p(x)$ is a mixture of distributions iff there exists $K, \pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K$, such that

$$p = \sum_{k=1}^{K} \pi_k p_{\theta_k}.$$

- Mixtures: convex combinations of distributions.
- Defines new parametric families of distributions. Parameter $\lambda = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ (for given K).

Mixtures and clustering

- Equivalence of mixture representation and existence of some hidden state variable Z with

$$\forall K \in \{1, \dots, K\},\ \pi_k = P(Z = k),\ \forall x\, p_{\theta_k}(x) = p(x|Z = k)$$

$$p(x) = \sum_{k=1}^{K} P(Z = k)p(x|Z = k) = \sum_{k=1}^{K} \pi_k p_{\theta_k}(x)$$

- Clustering: Z interpreted as the cluster of X (find Z).
- More generally: representation of heterogeneous sources (find λ).
- Potential use of mixtures in various settings (density estimation, ...).
- Cluster: essentially a conditional distribution – referred to as emission distribution – plus a prior on Z (see generative models for classification).

Remark

- Possible extensions to $p(.|z=k)$ being in different parametric families.
- Examples: mixtures of Weibull and Gamma distributions, mixtures of PDFs with respect to the same reference measures, mixtures of arbitrary measures, ...

$$\text{for } x \geq 0,\ p(x) =$$
$$0.2\frac{a}{b}\left(\frac{x}{b}\right)^{a-1}\exp\left[\left(-\frac{x}{b}\right)^{a}\right] + 0.8\frac{c^k}{(k-1)!}x^{k-1}\exp(-cx).$$

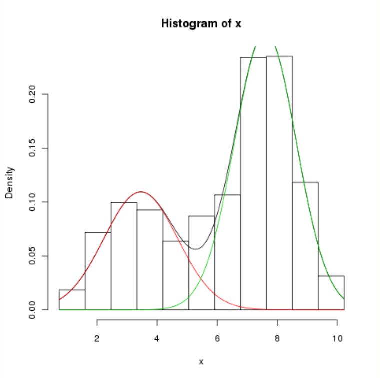
Interpretation: example (I)

- X: weight of some rodent.
- Proportion of females 1/3, males 2/3.
- Females generally lighter than males, the heaviest females potentially heavier than the lightest males.
- Gaussian (conditional) distribution of X for each gender. Dependence of means on gender.
- Unknown genders in population – both genders are mixed ("mixture").

Interpretation: example (II)

$$\pi_1 = 1/3;\ \pi_2 = 2/3;\ \mu_1 = 3;\ \mu_2 = 7;\ \sigma = 2$$

- If some rodent weighs 3g, its probability to be a female ($Z = 1$) is no longer 1/3 (compute it)



Histogram and mixture density function

Interpretation: example (III)

$$\pi_1 = 1/3;\ \pi_2 = 2/3;\ \mu_1 = 3;\ \mu_2 = 7;\ \sigma = 2$$

$$\begin{aligned} p_X(3) =& \frac{1}{3} \times \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{8}(3-3)^2\right) \\ &+ \frac{2}{3} \times \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{8}(7-3)^2\right) \end{aligned}$$

Hence,

$$P(Z=1|X=3) = \frac{p_X(3|Z=1)P(Z=1)}{p_X(3)}$$

$$= \frac{\frac{1}{3}\exp(-0)}{\frac{1}{3}\exp(-0) + \frac{2}{3}\exp(-2)}$$

$$\approx 0.79 > \frac{1}{3} = P(Z=1)$$

Identifiability issues

- Generally in statistics, a parametric family of models $\{p_\lambda\}_{\lambda \in \Lambda}$ has identifiable parameter iff $\forall (\lambda, \lambda') \in \Lambda^2,\ p_\lambda = p_{\lambda'} \Rightarrow \lambda = \lambda'$.
- Ensures uniqueness of parameter (interpretation, necessary condition for unique estimation, ...).
- Identifiability cannot be achieved for mixtures with variable and even fixed K, since for every $\lambda = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$, for every permutation $\kappa \in \mathcal{S}_K$ of the labels (values of the categorical value Z), if we set $\lambda' = (\pi_{\kappa(1)}, \ldots, \pi_{\kappa(K)}, \theta_{\kappa(1)}, \ldots, \theta_{\kappa(K)})$, we have

$$p_\lambda = \sum_{k=1}^{K} \pi_k p_{\theta_k} = \sum_{k=1}^{K} \pi_{\kappa(k)} p_{\theta_{\kappa(k)}} = p_{\lambda'}.$$

Identifiability for mixtures

- For mixture models: identifiability required up to a permutation of the labels (equivalence classes), requiring that

$$\sum_{k=1}^{K} \pi_k p_{\theta_k} = \sum_{k=1}^{K'} \pi'_k p'_{\theta_k}$$
$$\Rightarrow K = K' \text{ and } \exists \kappa \in \mathcal{S}_K \, \forall k,\ \pi_k = \pi'_{\kappa(k)} \text{ and } \theta_k = \theta'_{\kappa(k)}.$$

- Necessary condition: $\forall k,\ \pi_k > 0$.
- Additional sufficient condition for mixture identifiability: $\{p_\theta\}_{\theta \in \Theta}$ linearly independent PDFs.

Exercise 1

Show that the mixtures of uniform distributions $\{\mathcal{U}(a,b) | (a,b) \in \mathbb{R}^2, a < b\}$ are not identifiable.

Identifiability for mixtures (II)

Show that the mixtures of uniform distributions $\{\mathcal{U}(a,b) | (a,b) \in \mathbb{R}^2, a < b\}$ are not identifiable.
We consider

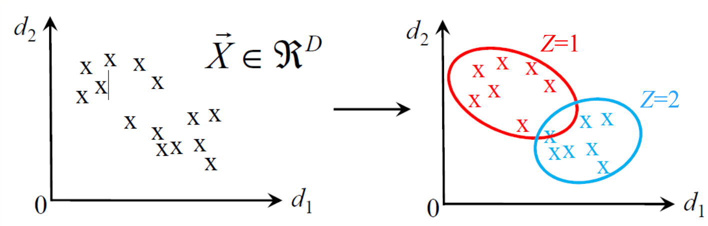
$$\frac{1}{2}\mathcal{U}(0,1) + \frac{1}{2}\mathcal{U}(1,2) = \mathcal{U}(0,2) = \frac{1}{4}\mathcal{U}(0,\frac{1}{2}) + \frac{3}{4}\mathcal{U}(\frac{1}{2},2).$$

Indeed,

$$\forall x \in [0,2],\ x \notin \{\frac{1}{2},1\} \Rightarrow \frac{1}{2}\mathbb{1}_{[0,1]}(x) + \frac{1}{2}\mathbb{1}_{[1,2]}(x) = \frac{1}{2}$$
$$= \mathbb{1}_{[0,2]}(x) = \frac{1}{4}\mathbb{1}_{[0,\frac{1}{2}]}(x) + \frac{3}{4}\mathbb{1}_{[\frac{1}{2},2]}(x).$$

Clustering with mixtures in three steps

1. $(Z_i, X_i)_{i=1,\dots,n}$ assumed independent (“independent mixture model”).
2. Parameter estimation (maximum likelihood) $\equiv$ learning from an unlabelled sample of size $n \ \rightarrow \hat{\lambda}_n$.
3. $\forall (i,k)$ compute $P_{\hat{\lambda}_n}(Z_i = k | X_i = x_i)$ (see previous slide)
4. MAP: $\forall i,\ \hat{Z}_i = \arg\max_k P_{\hat{\lambda}_n}(Z_i = k | X_i = x_i)$



Clustering with bivariate mixtures

Learning stage

$$\hat{\lambda}_n = \arg\max_{\lambda \in \mathcal{C}} \sum_{i=1}^{n} \ln\left[\sum_{k=1}^{K} \pi_k p_{\theta_k}(x_i)\right] = \arg\max_{\lambda \in \mathcal{C}} \ell_{x_1,\dots,x_n}(\lambda)$$

with $\mathcal{C} = \{(\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K) | \sum_k \pi_k = 1 \text{ and } \forall k\, \pi_k \geq 0\}$.

Likelihood equations:

$$\left.\begin{array}{l} \frac{\partial \ell_{x_1,\dots,x_n}}{\partial \pi_k}(\lambda) = \sum_{i=1}^{n} \frac{p_{\theta_k}(x_j)}{\sum_{l=1}^{K} \pi_l p_{\theta_l}(x_j)} = 0 \\ \nabla_{\theta_k} \ell_{x_1,\dots,x_n}(\lambda) = \sum_{i=1}^{n} \frac{\pi_k \nabla_{\theta_k} p_{\theta_k}(x_j)}{\sum_{l=1}^{K} \pi_l p_{\theta_l}(x_j)} = 0 \\ \text{and constraints...} \end{array}\right\}$$

- Non linear equations
- No closed form solution
- Newton method may work but what properties to expect from it?

EM algorithm: basic idea

Learning stage

Difficulty $\equiv$ non-linearity, from

$$\hat{\lambda}_n = \arg\max_{\lambda \in \mathcal{C}} \sum_{i=1}^{n} \underbrace{\ln \left[\sum_{k=1}^{K} \pi_k p_{\theta_k}(x_i) \right]}_{\text{Comes from } z \text{ being hidden}}$$

If known states $(Z_1, \ldots, Z_n)$: maximize

$$\sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \ln \left[\pi_k p_{\theta_k}(x_i)\right] = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \ln(\pi_k) + \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \ln p_{\theta_k}(x_i).$$

with same complexity as MLE estimation on K independent i.i.d. samples within family $\{p_\theta\}_{\theta \in \Theta}$.

EM algorithm: principle

$$\hat{\lambda} = \arg\max_{\lambda} \ln\left[p_\lambda(\boldsymbol{x})\right] = \arg\max_{\lambda} \ln\left[\sum_{\boldsymbol{z}} p_\lambda(\boldsymbol{x}, \boldsymbol{z})\right]$$

where:

- $\boldsymbol{x}$ observed, $\boldsymbol{z}$ hidden with finite values (arbitrary random vectors, extension to continuous $\boldsymbol{z}$ with integrals).
- $p_\lambda(\boldsymbol{x}, \boldsymbol{z})$ easy to maximize, $p_\lambda(\boldsymbol{x})$ difficult to maximize.

Example: $\boldsymbol{x} = (x_1, \ldots, x_n), \boldsymbol{z} = (z_1, \ldots, z_n)$

- Try to maximize $\ln p_\lambda(\boldsymbol{x}, \boldsymbol{z})$ rather than $\ln p_\lambda(\boldsymbol{x})$.
- Since $\boldsymbol{z}$ is hidden, consider $E[\ln p_\lambda(\boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \boldsymbol{x}]$.
- Expectation under which p_λ if λ and $\hat{\lambda}$ are unknown?
- $\arg\max_{\lambda} E_\lambda[\ln p_\lambda(\boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \boldsymbol{x}]$ seems too complicated: proceed iteratively.

EM algorithm: formulation

We need an initial value $\lambda^{(0)}$.

$$\lambda^{(m+1)} = \arg\max_{\lambda} E_{\lambda^{(m)}} \left[\ln p_{\lambda}(\boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \boldsymbol{x}\right] = \arg\max_{\lambda} Q(\lambda, \lambda^{(m)})$$

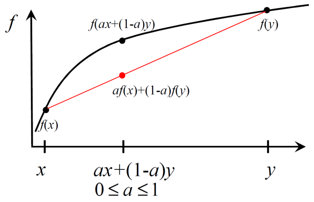
where

$$Q(\lambda, \lambda^{(m)}) = \sum_{\boldsymbol{z}} \ln p_{\lambda}(\boldsymbol{x}, \boldsymbol{z}) p_{\lambda^{(m)}}(\boldsymbol{z}|\boldsymbol{x}).$$

- E = Expectation (compute $E_{\lambda^{(m)}} \left[\ln p_{\lambda}(\boldsymbol{X}, \boldsymbol{Z}) | \boldsymbol{X} = \boldsymbol{x}\right]$ or at least any relevant quantity for the maximisation)
- M = maximisation (update parameter $\lambda \to \lambda^{(m+1)}$ using $\lambda^{(m)}$).

EM algorithm: main property

Theorem 1 (Dempster *et al.*, 1977).
$(\ln p_{\lambda^{(m)}}(x))_{m \geq 0}$ *is a non-decreasing sequence.*



Some concave function f

Remark 1.

- $(\ln p_{\lambda^{(m)}}(x))_{m \geq 0}$ *may not converge.*
- $(\lambda^{(m)})_{m \geq 0}$ *may not converge, or may converge to a saddle point, a local maximum, ...)*

Application of EM to clustering

Exercise 2

- Compute the so-called completed likelihood

$$\ell_{\boldsymbol{x},\boldsymbol{z}}(\lambda) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}_{\{z_i=k\}} \ln\left[\pi_k p_{\theta_k}(x_i)\right]$$

  and its maximizer $\hat{\lambda}_{\boldsymbol{x},\boldsymbol{z}}$ if $X \in \mathbb{R}^d$ has conditional multivariate Gaussian distribution with parameter $\theta = (\mu, \Sigma)$.
- We give

$$\nabla\mu\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right] = -2\Sigma^{-1}(x-\mu),$$
$$\nabla_\Sigma\left[(x-\mu)^T\Sigma^{-1}(x-\mu)\right] = -\Sigma^{-2}(x-\mu)(x-\mu)^T$$

and $\nabla_\Sigma\left[\ln(\text{det}(\Sigma))\right] = \Sigma^{-1}$.

EM algorithm: reestimation formulas

Exercise 3

- Read and answer the preparatory questions for next lab session (Independent mixture models).
- Give the reestimation formulas of the EM algorithm for independent mixtures with multivariate Gaussian emission distributions.

References

References

Dempster, A.P., Laird, N.M. and Rubin, D.B.
Maximum Likelihood from Incomplete Data via the EM Algorithm, (with discussion)
Journal of the Royal Statistical Society Series B **39**, pp. 1–38 (1977)

McLachlan, G.J. and Peel, D.
Finite Mixture Models
Wiley Series in Probability and Statistics, John Wiley and Sons, 2000.