Model Based Clustering

Fundamentals of Probabilistic Data Mining

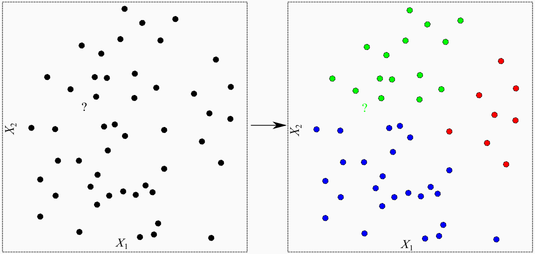
Fei Zheng

October, 2019

Clustering

Data: points $\{\boldsymbol{x}_n\}_{n=1}^N$ in $\mathbb{R}^d$
Aim: find K clusters (K fixed here)

- Distance-based approaches: close points tend to be in the same cluster. No explicit assumption required.



Clustering

Given a data set $\{\boldsymbol{x}_n\}_{n=1}^N$. Let z_n denote the cluster label of $\boldsymbol{x}_n$.

K-means (MacQueen 1967)

Objective: The sum of the squared distance of each $\boldsymbol{x}_n$ to its closest cluster centroid $\boldsymbol{\mu}_k$ is a minimum.

Repeat until converge:

- Assign $\boldsymbol{x}_n$ to the cluster indexed by the closest $\boldsymbol{\mu}_k$.
- Update $\boldsymbol{\mu}_k$ using the newly grouped data.

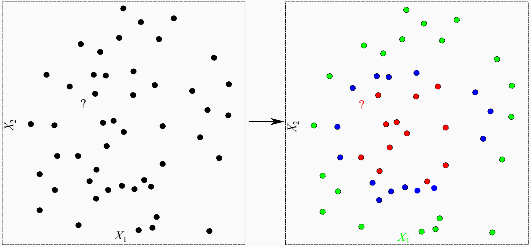
Clustering

Clustering vs Classification

	Clustering	Classification
Applied case	suggest groups based on patterns in data	classify new sample into known classes
Prior knowledge	No prior knowledge	A training set
Data needs	Unlabeled samples	Labeled samples from knowing classes

Clustering

- Model-based approaches: If $z_i = z_j = k$, then $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ should have the same (conditional) distribution p_k.



Mixture Models

Definition (McLachalan & Peel, 2000)

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

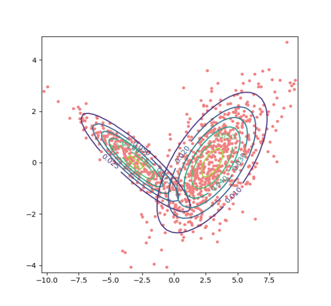
$$p = \sum_{k=1}^{K} \pi_k p_{\theta_k}, \qquad 0 \leq \pi_k \leq 1, \ \sum_{k=1}^{K} \pi_k = 1$$

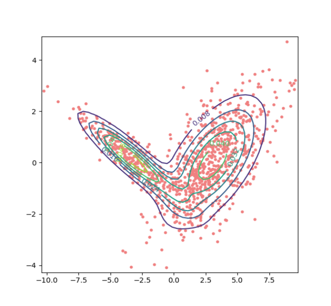
- Mixture model is convex combination of distributions .
- Defines new parametric families of distributions.
  Parameters: $\Theta = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K)$ (for given K).

Mixture Models

Example: 2-component Gaussian Mixture Model (GMM)

- $K = 2$; $\pi_1 = 0.3, \pi_2 = 0.7$; $p_{\theta_1} = \mathcal{N}\left(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1\right), p_{\theta_2} = \mathcal{N}\left(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2\right)$.
- $p(\boldsymbol{x}) = \pi_1 p_{\theta_1}(\boldsymbol{x}) + \pi_2 p_{\theta_2}(\boldsymbol{x})$.





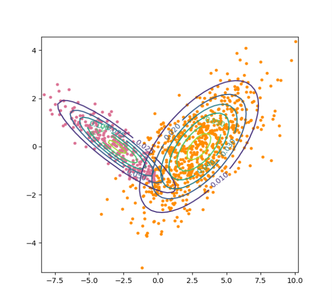
Mixture Model and Clustering

Equivalence of mixture representation and existence of some discrete hidden random variable Z (label of cluster):

$$\forall k \in \{1, \dots, K\}, \pi_k = p(Z = k)$$

$$\forall \boldsymbol{x},\ p_{\theta_k}(\boldsymbol{x}) = p(\boldsymbol{x}|Z = k)$$

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(Z = k) p(\boldsymbol{x}|Z = k)$$



- **Clustering**: to infer Z interpreted as the cluster label (heterogeneous sources) given $\boldsymbol{x}$.

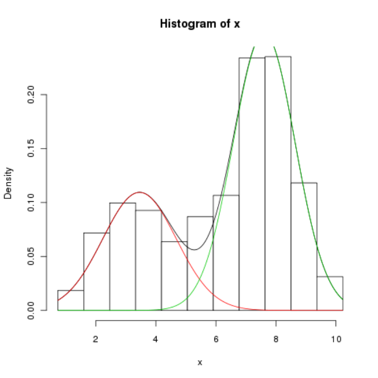
Interpretation

- 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.
- The weights (X) are (conditionally) Gaussian distributed for each gender, depend on mean weight and variance.
- Unknown genders in population – both genders are mixed ("mixture of 2 Gaussians").

Interpretation

- Parameters:
  $\pi_1 = 1/3$, $\pi_2 = 2/3$;
  $\mu_1 = 3$ (kg), $\mu_2 = 7$ (kg);
  $\sigma_1 = \sigma_2 = 2$.

- If some rodent weighs 3 (kg), what is its probability to be a female $(Z = 1)$?



Remark

- Possible extensions to $p(.|z = k)$ being in different parametric families.
- Example: mixtures of Weibull and Gamma distributions.

$$\text{for } x \geq 0,\ p(x) =$$

$$0.2 \frac{a}{b} \left(\frac{x}{b}\right)^{a-1} \exp\left[\left(-\frac{x}{b}\right)^{a}\right] + 0.8 \frac{c^k}{(k-1)!} x^{k-1} \exp(-cx).$$

Identifiability Issues

- Generally, a parametric family of models $\{p_\Theta\}_{\Theta \in \boldsymbol{\Theta}}$ has identifiable parameter iff

$$\forall (\Theta, \Theta') \in \boldsymbol{\Theta}^2,\ p_\Theta = p_{\Theta'} \Rightarrow \Theta = \Theta'.$$

- Ensures uniqueness of parameter (necessary condition for unique estimation).
- Identifiability cannot be achieved for mixtures even with fixed K, since for true $\Theta = (\pi_1, \dots, \pi_K, \theta_1, \dots, \theta_K)$, for any permutation $\kappa \in \mathcal{S}_K$ of the labels (categorical value Z), if set $\Theta' = (\pi_{\kappa(1)}, \dots, \pi_{\kappa(K)}, \theta_{\kappa(1)}, \dots, \theta_{\kappa(K)})$, we have

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

Identifiability for Mixtures

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

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

- To ensure this, we constraint the mixture models to satisfy $\forall k,\, \pi_k > 0$ (negative case: Zhang & Zhang).
- Additional sufficient condition for mixture identifiability: $\{p_\theta\}_{\theta \in \boldsymbol{\theta}}$ are linearly independent PDFs.

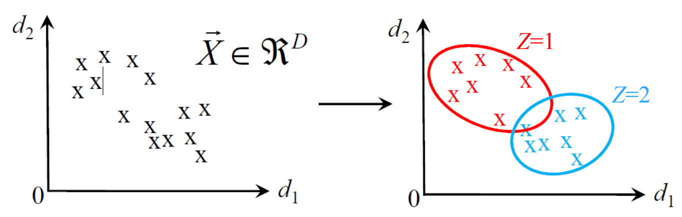
Exercise 1

Proof that the mixtures of uniform distributions are not identifiable.

Clustering with Mixtures in 3 Steps

Assumption: $\{Z_n, X_n\}_{n=1}^{N}$ are independent. The families of p_{θ_k} are known (good candidates).

1. Parameter estimation through Maximum Likelihood: learning $\hat{\Theta}$ from an unlabelled sample set of size N.
2. $\forall (n, k)$, compute $p_{\hat{\Theta}}(Z_n = k | X_n = x_n)$.
3. MAP: $\forall n,\ \hat{Z}_n = \arg\max_k p_{\hat{\Theta}}(Z_n = k | X_n = x_n)$.



Learning Stage

Maximum Likelihood Estimation (MLE):

$$\hat{\Theta} = \arg\max_{\Theta \in \mathcal{C}} \ell_{\boldsymbol{x}}(\Theta) = \arg\max_{\Theta \in \mathcal{C}} \sum_{n=1}^{N} \underbrace{\ln \left[\sum_{k=1}^{K} \pi_k p_{\theta_k}(x_n) \right]}_{z \text{ is hidden inside}}$$

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

Set derivatives to 0:

$$\frac{\partial \ell_{\boldsymbol{x}}}{\partial \pi_k}(\Theta) = 0, \quad \nabla_{\theta_k} \ell_{\boldsymbol{x}}(\Theta) = 0$$

No close form solution...

Learning Stage

$$\hat{\Theta} = \arg\max_{\Theta \in \mathcal{C}} \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k p_{\theta_k}(x_n) \right]$$

If the hidden states $(z_1, \ldots, z_N)$ are known:

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

with the same complexity as to MLE estimation K different distributions within family $\{p_\theta\}_{\theta \in \boldsymbol{\theta}}$ on K independent sample sets.

EM Algorithm: Principle

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

where:

- $\boldsymbol{x} = \{x_1, \cdots, x_N\}$ observed, $\boldsymbol{z} = \{z_1, \cdots, z_N\}$ hidden discrete finite values labels.
- $p_{\Theta}(\boldsymbol{x}, \boldsymbol{z})$ easy to maximize, $p_{\Theta}(\boldsymbol{x})$ difficult to maximize.

Principle:

- Maximize $\ln p_{\Theta}(\boldsymbol{x}, \boldsymbol{z})$ instead of $\ln p_{\Theta}(\boldsymbol{x})$.
- Since $\boldsymbol{z}$ is hidden, consider $\mathrm{E}_{(\boldsymbol{Z}|\boldsymbol{x},\Theta)}[\ln p_{\Theta}(\boldsymbol{X}, \boldsymbol{Z})|\boldsymbol{X} = \boldsymbol{x}]$.
- Maximize $\sum_{\boldsymbol{z}} p(\boldsymbol{z}|\boldsymbol{x}, \Theta) \ln p_{\Theta}(\boldsymbol{x}, \boldsymbol{z})$ iteratively.

EM Algorithm: Formulation

1. Initialize $\Theta^{(i=0)}$.
2. E-step (expectation): compute $\mathcal{Q}\left(\Theta, \Theta^{(i)}\right)$ that

$$\mathcal{Q}\left(\Theta, \Theta^{(i)}\right) = \sum_{\boldsymbol{z}} p\left(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(i)}\right) \ln p_{\Theta}(\boldsymbol{x}, \boldsymbol{z})$$

   or at least any relevant quantity.
3. M-step (maximization): update Θ by

$$\Theta^{(i+1)} = \arg\max_{\Theta} \mathcal{Q}\left(\Theta, \Theta^{(i)}\right)$$

4. Check the convergence of the log-likelihood $\ln p_{\Theta}(\boldsymbol{x})$.
   Back to step 2 if convergence criterion is not satisfied.

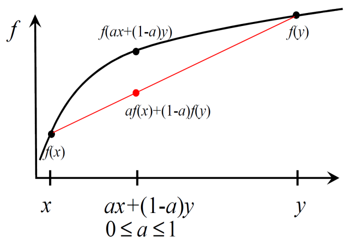
EM Algorithm: Main Property

Theorem (Dempster *et al.*, 1977)

$(\ln p_{\Theta^{(i)}}(x))_{i \geq 0}$ *is a non-decreasing sequence.*

$$\ln\left[\sum_z p_\Theta(\mathbf{x},\mathbf{z})\right] = \ln\left[\sum_z q(\mathbf{z})\frac{p_\Theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right] \geq \underbrace{\sum_z q(\mathbf{z})\ln\left[\frac{p_\Theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})}\right]}_{\mathcal{L}(q,\Theta)}$$

with equality when $\frac{p_\Theta(\mathbf{x},\mathbf{z})}{q(\mathbf{z})} = C$

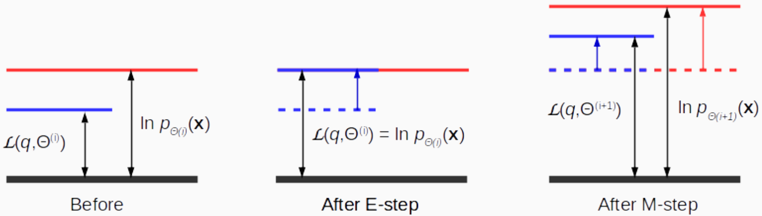


Some concave function $f(\cdot)$

EM Algorithm: Main Property

E-step: adjust $q(\boldsymbol{z}) = p\left(\boldsymbol{z}|\boldsymbol{x}, \Theta^{(i)}\right)$ to make $\ln p_{\Theta^{(i)}}(\boldsymbol{x}) = \mathcal{L}\left(q, \Theta^{(i)}\right)$

M-step: adjust Θ (derivative $= 0$) to maximize $\mathcal{L}\left(q, \Theta^{(i)}\right)$



- **Remark**: $\ln p_{\Theta^{(i)}}(\boldsymbol{x})$ may converge to saddle point, local maximum or even not converge.

Application

Mixtures of Bernoulli distributions

Consider a set of M binary variables $\boldsymbol{x} = (x_1, \cdots, x_M)^\intercal$ follows

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{\mu}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k) \text{ where } p(\boldsymbol{x}|\boldsymbol{\mu}_k) = \prod_{m=1}^{M} \mu_{km}^{x_m}(1-\mu_{km})^{1-x_m}$$

- To cluster a given data set $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$:
  1. Estimate the parameters of the clusters $\hat{\Theta} = \{\hat{\boldsymbol{\pi}}, \hat{\boldsymbol{\mu}}\}$

     E-step: $\gamma^{(i)}(z_{nk}) = p(Z_n = k|\boldsymbol{x}_n, \Theta^{(i)}) = \frac{\pi_k^{(i)} p(\boldsymbol{x}_n|\boldsymbol{\mu}_k^{(i)})}{\sum_{l=1}^{K} \pi_l^{(i)} p(\boldsymbol{x}_n|\boldsymbol{\mu}_l^{(i)})}$.

     M-step: $\pi_k^{(i+1)} = \frac{N_k^{(i)}}{N}$, $\boldsymbol{\mu}_k^{(i+1)} = \frac{1}{N_k^{(i)}} \sum_{n=1}^{N} \gamma^{(i)}(z_{nk})\boldsymbol{x}_n$ with $N_k^{(i)} = \sum_{n=1}^{N} \gamma^{(i)}(z_{nk})$.
  2. Estimate label $\hat{Z}_n = \arg\max p(Z_n = k|\boldsymbol{x}_n, \hat{\Theta})$

Exercise 2

Compute the so-called complete-data log-likelihood:

$$\ell_{\boldsymbol{x},\boldsymbol{z}}(\Theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{1}_{\{z_n=k\}} \ln\left[\pi_k p_{\theta_k}(x_n)\right]$$

and its maximizer $\hat{\Theta}_{\boldsymbol{x},\boldsymbol{z}}$ if $X \in \mathbb{R}^d$ has conditional multivariate Gaussian distribution that $p_{\theta_k}(X_n|Z_n = k) = \mathcal{N}(\mu_k, \Sigma_k)$.

We have:

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

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

Exercise 3

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

References i

[1] Arthur P Dempster, Nan M Laird, and Donald B Rubin.
Maximum likelihood from incomplete data via the em algorithm.
Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22, 1977.

[2] James MacQueen et al.
Some methods for classification and analysis of multivariate observations.
In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[3] Geoffrey McLachlan and David Peel.
Finite mixture models.
John Wiley & Sons, 2004.

References ii

[4] Baibo Zhang and Changshui Zhang.
Finite mixture models with negative components.
In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 31–41. Springer, 2005.