Machine Learning, Lecture 4: Gaussian Mixture Model & EM algorithm

S. Nõmm

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

26.02.2015

Latent Variable Models

Latent Variable Models (**LVM**) - models with hidden variables.
An important assumption is that observed variables are correlated because they arise from a hidden common "cause". Let $z_{i,1}, \dots, z_{i,L}$ are L latent variables, and $x_{i,1}, \dots, x_{i,D}$ are D visible variables.
The form of the likelihood $\mathcal{L}(x_i \mid z_i)$ and the prior $p(z_i)$ defines the model.

Variety of LVMs

The form of the likelihood $p(x_i \mid z_i)$ and the prior $p(z_i)$ lead following models

$p(x_i \mid z_i)$	$p(z_i)$	Name
MVN	Discr.	Mixture of Gaussians
Prod. Discr.	Discr.	Mixture of Multinominals
Prod. Gauss.	Prod. Gauss.	Factor analysis/probabilitstic PCA
Prod. Gauss.	Prod. Laplace	Probabilistic ICA/sprase coding
Prod. Discr.	Prod. Gauss.	Multinominal PCA
Prod. Gauss.	Dirichlet	Latent Dirichlet allocation
Prod. Noisy-QR.	Prod.Bernoulli	BN20/QMR
Prod. Bernoulli.	Prod. Bernoulli	Sigmoid belief net

Mixture models

Let $z_i = \{1, \ldots, K\}$, - discrete latent states.

$$\begin{aligned} p(z_i) &= \mathrm{Cat}(\pi) \\ \mathcal{L}(x_i \mid z_i = k) &= p_k(x_i) \end{aligned}$$

Overall model is known as *Mixture model* (we are mixing together K base distributions)

$$p(x_i \mid \theta) = \sum_{k=1}^{K} \pi_k p_k(x_i \mid \theta)$$

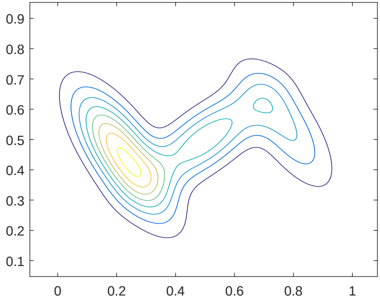
where mixed weights π_k satisfy $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$

Mixture of Gaussians

Mixture of Gaussian (MOG) is the most widely used mixture model. Each base distribution is a multivariate Gaussian with mean μ_k and covariance matrix Σ_k

$$p(x_i \mid \theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i \mid \mu_k, \Sigma_k)$$

Mixture of Gaussians



Mixture of Gaussians

- Latent variables z_i : $\quad z_i = k$ component k generated point x_i.
- $p(z_i = k \mid \pi) = \pi_k$ - probability of being generated by a component.
- $p\big(\boldsymbol{x_i} \mid z_i = k, \boldsymbol{\mu}, \boldsymbol{\Sigma}\big) = \mathcal{N}\big(\boldsymbol{x_i} \mid \boldsymbol{\mu}_k, \sigma_k\big)$ - probability of a given point whereas it is known which component generated it.
- $p\big(\boldsymbol{x_i}, z_i = k \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\big) = \pi_k \mathcal{N}\big(\boldsymbol{x_i} \mid \boldsymbol{\mu}_k, \Sigma_k\big)$ - joint probability of generating the component and the point from it.
- $p\big(\boldsymbol{x_i} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\big) = \sum_{k=1}^{K} \pi_k \mathcal{N}\big(\boldsymbol{x_i} \mid \boldsymbol{\mu}_k, \Sigma_k\big)$ - *marginal probability* of the point.

Parameter estimation for Gaussian Mixture Models

- The goal is to estimate parameters: $\boldsymbol{\pi}, \boldsymbol{\mu_k}, \Sigma_k, \quad k = 1, \ldots, K$
- The log-likelihood function of GMM is

$$\log p\big(\boldsymbol{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\big) = \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \Sigma_k)\Big)$$

- Possible problems:
  - Unidentifiability: K-component mixture has $K!$ possible labeling therefore there is no unique maximal likelihood estimate and in turn no unique maximum a posterior estimate.
  - Summation inside the logarithm ... .

Observe the following

- The knowledge of component parameters and mixing proportions allows to compute the probability that the component k responsible [1] for the i-th point $p(z_i = k \mid \boldsymbol{x_i}, \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The knowledge of the responsibilities allows to compute the estimates for the mixing coefficients π_k.
- The knowledge of responsibilities and mixing coefficients allows to compute the estimates for component means μ_k and variances Σ_k

This leads the idea of two step iterative algorithm:

- **Step E:** Inferring the missing values given the parameters.
- **Step M:** Optimization of the parameters given the "filled data".

---

[1] Responsibility of the cluster k for point i is the posterior probability that point i belongs to cluster k, $p(z_i = k \mid \boldsymbol{x_i}, \boldsymbol{\theta})$

Expectation - Maximization

Expectation - Maximization (EM):

- Let x_i denote the visible observed values in case i, and z_i - hidden or missing variables. The goal is to maximize the log likelihood of the observed data:

$$\mathcal{L}(\theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta) = \sum_{i=1}^{N} \log \Big[\sum_{z_i} p(x_i, z_i \mid \theta) \Big]$$

- Way around the problem with the sum under the log. Define the complete data log likelihood as is follows

$$\mathcal{L}_c(\theta) = \sum_{i=1}^{N} \log p(x_i, z_i \mid \theta)$$

Note, that this could not be computed due to the fact that z_i are unknown.

EM

- Define expected complete data log likelihood:

$$Q(\theta, \theta^{t-1}) = \mathbb{E}[l_c(\theta) \mid \mathcal{D}, \theta^{t-1}].$$

  here t is the iteration number. Q will be referred as *auxiliary function*.
- **E** step computes the latent values needed to compute $Q(\theta \mid \theta^{t-1})$.
- **M** step optimizes Q with respect to θ.

$$\theta^t = \arg\max_\theta Q(\theta, \theta^{t-1})$$

EM -algorithm

- Auxiliary function:

$$Q(\theta, \theta^{t-1}) = \sum_i \sum_k r_{i,k} \log \pi_k + \sum_i \sum_k r_{i,k} \log p(\boldsymbol{x_i} \mid \theta_k).$$

- **E step:** compute the responsibilities $r_{i,k}$ for each i and k:

$$r_{i,k} = \frac{\pi_k p(\boldsymbol{x_i} \mid \theta_k^{t-1})}{\sum_{k'} \pi_{k'} p(\boldsymbol{x_i} \mid \boldsymbol{\theta}_{k'}^{t-1})}.$$

EM -algorithm

- Optimize Q with respect to $\boldsymbol{\pi}, \boldsymbol{\mu}_k, \Sigma_k$.
- $$\pi_k = \frac{1}{N}\sum_i r_{i,k} = \frac{r_k}{N}$$

  where $r_k = \sum_i r_{i,k}$
- Derive **M step** for the μ_k and Σ_k

$$\mathcal{L}(\mu_k, \Sigma_k) = -\frac{1}{2}\sum_i r_{i,k}[\log \mid \Sigma_k \mid + (x_i - \mu_k)^T \sigma_k^{-1}(x_i - \mu k)]$$

$$\begin{aligned}
\mu_k &= \frac{\sum_i r_{i,k} x_i}{r_k} \\
\Sigma_k &= \frac{\sum_i r_{i,k} x_i x_i^t}{r_k} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T
\end{aligned}$$

EM & ?

-

$$\mu_k = \frac{\sum_i r_{i,k} x_i}{r_k}$$

$$\Sigma_k = \frac{\sum_i r_{i,k} x_i x_i^t}{r_k} - \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T$$

- Let us suppose now that all the covariances are set to the same symmetric matrix for each cluster.

$$\Sigma_1 = \ldots = \Sigma_K = \sigma^2 \boldsymbol{I}$$

- Let us further suppose that mixing properties are uniform $\pi_k = frac 1 K$
- The only parameter to estimate are cluster means μ_k
- We got ?