Pattern Recognition and Machine Learning
Chapter 9: Mixture Models and EM

Thomas Mensink   Jakob Verbeek

October 11, 2007

Le Menu

Mixture Models and EM: Introduction

- Additional latent variables allows to express relatively complex marginal distributions over latent variables in terms of more tractable joint distributions over the expanded space.
- Maximum-Likelihood estimator in such a space is the *Expectation-Maximization* (EM) algorithm.
- Chapter 10 provides Bayesian treatment using variational inference

K-Means Clustering: Distortion Measure

- Dataset $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
- Partition in K clusters

- Cluster prototype: μ_k
- Binary indicator variable, 1-of-K Coding scheme
  $r_{nk} \in \{0, 1\}$
  $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$.
  Hard assignment.
- Distortion measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \quad (9.1)$$

K-Means Clustering: Expectation Maximization

- Find values for $\{r_{nk}\}$ and $\{\mu_k\}$ to minimize:

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \tag{9.1}$$

- Iterative procedure:
  1. Minimize J w.r.t. r_{nk}, keep μ_k fixed (Expectation)

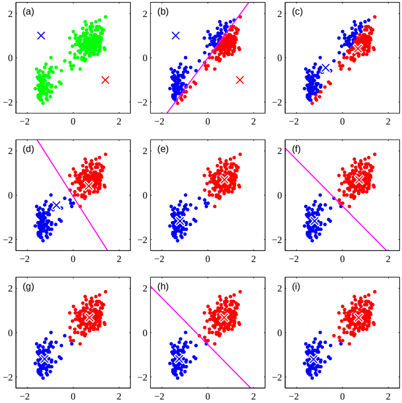
$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_k\|^2 \\ 0 & \text{otherwise} \end{cases} \tag{9.2}$$

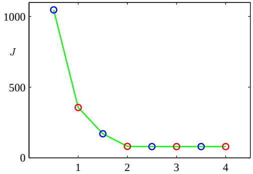
  2. Minimize J w.r.t. μ_k, keep r_{nk} fixed (Maximization)

$$2 \sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0 \tag{9.3}$$

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}} \tag{9.4}$$

K-Means Clustering: Example





- Each E or M step reduces the value of the objective function J
- Convergence to a global or local maximum

K-Means Clustering: Concluding remarks

1. Direct implementation of K-Means can be slow
2. Online version:

$$\mu_k^{\mathrm{new}} = \mu_k^{\mathrm{old}} + \eta_n(\mathbf{x}_n - \mu_k^{\mathrm{old}}) \tag{9.5}$$

3. K-mediods, general distortion measure

$$\tilde{J} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \mathcal{V}(\mathbf{x}_n, \mu_k) \tag{9.6}$$

   where $\mathcal{V}(\cdot, \cdot)$ is any kind of dissimilarity measure

4. Image segmentation and compression example:

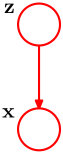
4.2 % | 8.3 % | 16.7 % | 100 %

Mixture of Gaussians: Latent variables

- Gaussian Mixture Distribution:

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \tag{9.7}$$

- Introduce latent variable $\mathbf{z}$
    - $\mathbf{z}$ is binary 1-of-K coding variable
    - $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$



Mixture of Gaussians: Latent variables (2)

- $p(z_k = 1) = \pi_k$
  constraints: $0 \leq \pi_k \leq 1$, and $\sum_k \pi_k = 1$
  $p(\mathbf{z}) = \prod_k \pi_k^{z_k}$

- $p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$
  $p(\mathbf{x}|\mathbf{z}) = \prod_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k}$

- $p(\mathbf{x}) = \sum_z p(\mathbf{x}, \mathbf{z}) = \sum_z p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_k \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$

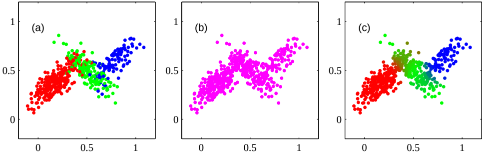
- The use of the joint probability $p(\mathbf{x}, \mathbf{z})$, leads to significant simplifications

Mixture of Gaussians: Latent variables (3)

- responsibility of component k to generate observation $\mathbf{x}$ (9.13):

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_k p(z_k = 1)p(\mathbf{x}|z_k = 1)}$$
$$= \frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}$$

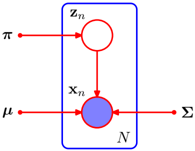
  is the *posterior probability*
- Generate random samples with ancestral sampling:
  First generate $\hat{\mathbf{z}}$ from $p(\mathbf{z})$
  Second generate a value for $\mathbf{x}$ from $p(\mathbf{x}|\hat{\mathbf{z}})$
  See Chapter 11.



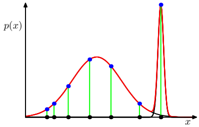
Mixture of Gaussians: Maximum Likelihood

- Log Likelihood

$$\ln p(\mathbf{X}|\pi,\mu,\Sigma) = \sum_{n=1}^{N} \ln\left\{\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k,\Sigma_k)\right\} \quad (9.14)$$



- Singularity when a mixture component collapses on a datapoint
- Identifiability for a ML solution in a K-component mixture there are $K!$ equivalent solutions.



Mixture of Gaussians: EM for Gaussian Mixtures

- Informal introduction of *expectation-maximization* algorithm (Dempster *et al.*, 1977).
- Maximum of log likelihood:
  derivatives of $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ w.r.t parameters to 0.

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \right\} \tag{9.14}$$

- For the μ_k[1]:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}}_{\gamma(z_k)} \Sigma_k^{-1}(\mathbf{x}_n - \mu_k) \tag{9.16}$$

$$\mu_k = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)\mathbf{x}_n \tag{9.17}$$

[1] Error in book, see erata file

Mixture of Gaussians: EM for Gaussian Mixtures

- For Σ_k:

$$\Sigma_k = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \tag{9.19}$$

- For the π_k:
  - Take into account constraint $\sum_k \pi_k = 1$
  - Lagrange multiplier

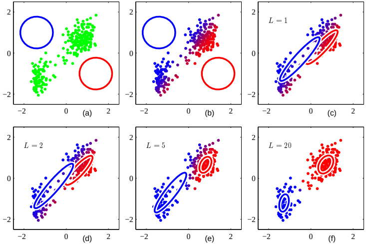
$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda(\sum_k \pi_k - 1) \tag{9.20}$$

$$0 = \sum_n \frac{\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)} + \lambda \tag{9.21}$$

$$\pi_k = \frac{\sum_n \gamma(z_k)}{N} \tag{9.22}$$

Mixture of Gaussians: EM for Gaussian Mixtures Example

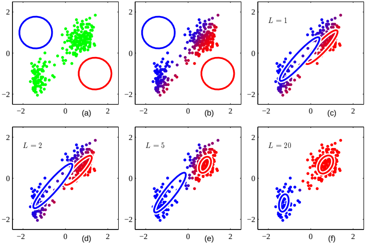
- No closed form solutions: $\gamma(z_k)$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood:
  Alternate between estimating the current $\gamma(z_k)$ and updating the parameters $\{\mu_k, \Sigma_k, \pi_k\}$.



- More iterations needed to converge than K-means algorithm, and each cycle requires more computation
- Common, initialise parameters based K-means run.

Mixture of Gaussians: EM for Gaussian Mixtures Example

- No closed form solutions: $\gamma(z_k)$ depends on parameters
- But these equations suggest simple iterative scheme for finding maximum likelihood:
  Alternate between estimating the current $\gamma(z_k)$ and updating the parameters $\{\mu_k, \Sigma_k, \pi_k\}$.



- More iterations needed to converge than K-means algorithm, and each cycle requires more computation
- Common, initialise parameters based K-means run.

Mixture of Gaussians: EM for Gaussian Mixtures Summary

1. Initialize $\{\mu_k, \Sigma_k, \pi_k\}$ and evaluate log-likelihood
2. E-Step Evaluate responsibilities $\gamma(z_k)$
3. M-Step Re-estimate paramters, using current responsibilities:

$$\mu_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k) \mathbf{x}_n \tag{9.23}$$

$$\Sigma_k^{\text{new}} = \frac{1}{\sum_n \gamma(z_k)} \sum_n \gamma(z_k)(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \tag{9.24}$$

$$\pi_k^{\text{new}} = \frac{\sum_n \gamma(z_k)}{N} \tag{9.25}$$

4. Evaluate log-likelihood $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ and check for convergence (go to step 2).

An Alternative View of EM: latent variables

- Let $\mathbf{X}$ observed data, $\mathbf{Z}$ latent variables, θ parameters.
- Goal: maximize marginal log-likelihood of observed data

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}. \tag{9.29}$$

- Optimization problematic due to log-sum.
- Assume straightforward maximization for complete data

$$\ln p(\mathbf{X}, \mathbf{Z}|\theta).$$

- Latent $\mathbf{Z}$ is known only through $p(\mathbf{Z}|\mathbf{X}, \theta)$.
- We will consider expectation of complete data log-likelihood.

An Alternative View of EM: algorithm

- **Initialization:** Choose initial set of parameters θ^{old}.
- **E-step:** use current parameters θ^{old} to compute $p(\mathbf{Z}|\mathbf{X}, \theta^{old})$ to find expected complete-data log-likelihood for general θ

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}|\theta). \qquad (9.30)$$

- **M-step:** determine θ^{new} by maximizing (9.30)

$$\theta^{new} = \arg\max_{\theta} \mathcal{Q}(\theta, \theta^{old}). \qquad (9.31)$$

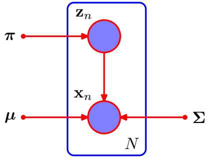
- **Check convergence:** stop, or $\theta^{old} \leftarrow \theta^{new}$ and go to **E-step**.

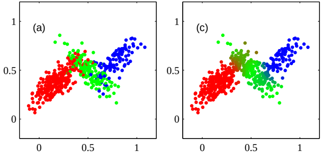
An Alternative View of EM: Gaussian mixtures revisited

- For mixture assign each $\mathbf{x}$ latent assigment variables z_k.

- Complete-data (log-)likelihood (9.36), and expectation (9.40)

$$
\begin{aligned}
p(\mathbf{x}, \mathbf{z}|\theta) &= \prod_{k=1}^{K} \pi_k^{z_k} \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)^{z_k} \\
\ln p(\mathbf{x}, \mathbf{z}|\theta) &= \sum_{k=1}^{K} z_k \left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\} \\
\mathcal{Q}(\theta) = \mathbb{E}_{\mathbf{z}}[\ln p(\mathbf{x}, \mathbf{z}|\theta)] &= \sum_{k=1}^{K} \gamma(z_k) \left\{\ln \pi_k + \ln \mathcal{N}\left(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right\}
\end{aligned}
$$





Example EM algorithm: Bernoulli mixtures

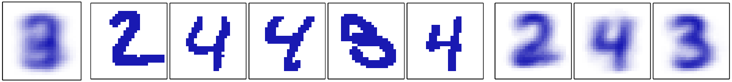
- Bernoulli distributions over binary data vectors

$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i}(1-\mu_i)^{(1-x_i)}. \qquad (9.44)$$

- Mixture of Bernoullis can model variable correlations.

- As the Gaussian, Bernoulli is member of exponential family
  - model log-linear, mixture not, complete-data log-likelihood is.

- Simple EM algorithm to find ML parameters
  - **E-step:** compute responsibilities $\gamma(z_{nk}) \propto \pi_k p(\mathbf{x}_n|\boldsymbol{\mu}_k)$
  - **M-step:** update parameters $\pi_k = N^{-1}\sum_n \gamma(z_{nk})$, and $\boldsymbol{\mu}_k = (N\pi_k)^{-1}\sum_n \gamma(z_{nk})\mathbf{x}_n$



Example EM algorithm: Bayesian linear regression

- Recall Bayesian linear regression: it's a latent variable model

$$p(\mathbf{t}|\mathbf{w},\beta,\mathbf{X}) = \prod_{n=1}^{N} \mathcal{N}\left(t_n; \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\right), \quad (3.10)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}\left(\mathbf{w}; 0, \alpha^{-1}\mathbf{I}\right), \quad (3.52)$$

$$p(\mathbf{t}|\alpha,\beta,\mathbf{X}) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}. \quad (3.77)$$

- Simple EM algorithm to find ML parameters (α, β)
  - **E-step:** compute responsibilities over latent variable $\mathbf{w}$

$$p(\mathbf{w}|\mathbf{t},\alpha,\beta) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{S}), \quad \mathbf{m} = \beta\mathbf{S}\boldsymbol{\Phi}^\top\mathbf{t}, \quad \mathbf{S}^{-1} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi}.$$

  - **M-step:** update parameters using complete-data log-likelihood

$$\alpha^{-1} = (1/M)\left(\mathbf{m}^\top\mathbf{m} + \mathrm{Tr}\{\mathbf{S}\}\right), \quad (9.63)$$

$$\beta^{-1} = (1/N)\sum_{n=1}^{N}\left\{t_n - \mathbf{m}^\top\phi(\mathbf{x}_n)\right\}^2.$$

The EM Algorithm in General

- Let $\mathbf{X}$ observed data, $\mathbf{Z}$ latent variables, θ parameters.
- Goal: maximize marginal log-likelihood of observed data

$$\ln p(\mathbf{X}|\theta) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \right\}. \tag{9.29}$$

- Maximization of $p(\mathbf{X}, \mathbf{Z}|\theta)$ simple, but difficult for $p(\mathbf{X}|\theta)$.
- Given any $q(\mathbf{Z})$, we decompose the data log-likelihood

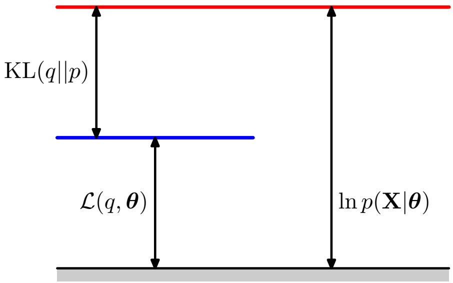
$$\begin{aligned}
\ln p(\mathbf{X}|\theta) &= \mathcal{L}(q, \theta) + \mathrm{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta)), \\
\mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})}, \\
\mathrm{KL}(q(\mathbf{Z}) \| p(\mathbf{Z}|\mathbf{X}, \theta)) &= -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \geq 0.
\end{aligned}$$

The EM Algorithm in General: the EM bound

- $\mathcal{L}(q, \theta)$ **is a lower bound on the data log-likelihood**
  - $-\mathcal{L}(q, \theta)$ known as variational free-energy

$$\mathcal{L}(q, \theta) = \ln p(\mathbf{X}|\theta) - \mathrm{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \theta) \leq \ln p(\mathbf{X}|\theta)$$

- **The EM alogorithm performs coordinate ascent on $\mathcal{L}$**
  - E-step maximizes $\mathcal{L}$ w.r.t. q for fixed θ
  - M-step maximizes $\mathcal{L}$ w.r.t. θ for fixed q

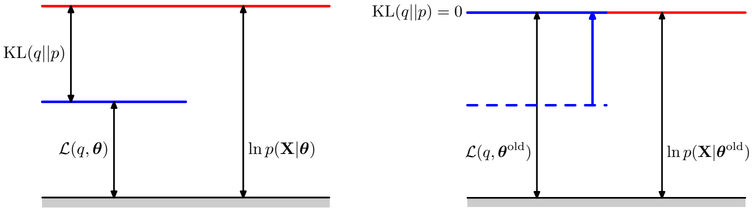


The EM Algorithm in General: the E-step

- E-step maximizes $\mathcal{L}(q,\theta)$ w.r.t. q for fixed θ

$$\mathcal{L}(q,\theta) \quad = \quad \ln p(\mathbf{X}|\theta) - \mathrm{KL}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{X},\theta))$$

- $\mathcal{L}$ maximized for $q(\mathbf{Z}) \leftarrow p(\mathbf{Z}|\mathbf{X},\theta)$

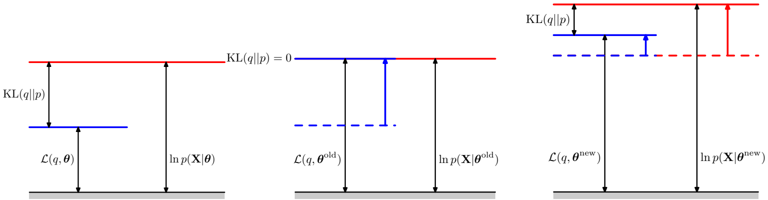


The EM Algorithm in General: the M-step

- M-step maximizes $\mathcal{L}(q, \theta)$ w.r.t. θ for fixed q

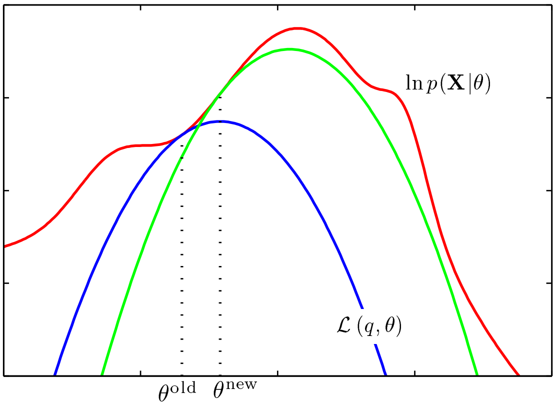
$$\mathcal{L}(q, \theta) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z})$$

- $\mathcal{L}$ maximized for $\theta = \arg\max_{\theta} \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z}|\theta)$



The EM Algorithm in General: picture in parameter space

- E-step resets bound $\mathcal{L}(q, \theta)$ on $\ln p(\mathbf{X}|\theta)$ at $\theta = \theta^{old}$, it is
  - tight at $\theta = \theta^{old}$
  - tangential at $\theta = \theta^{old}$
  - convex (easy) in θ for exponential family mixture components



The EM Algorithm in General: Final Thoughts

- (local) maxima of $\mathcal{L}(q, \theta)$ correspond to those of $\ln p(\mathbf{X}|\theta)$

- EM converges to (local) maximum of likelihood
  - Coordinate ascent on $\mathcal{L}(q, \theta)$, and $\mathcal{L} = \ln p(\mathbf{X}|\theta)$ after E-step

- Alternative schemes to optimize the bound
  - Generalized EM: relax M-step from maxmizing to increasing $\mathcal{L}$
  - Expectation Conditional Maximization: M-step maximizes w.r.t. groups of parameters in turn
  - Incremental EM: E-step per data point, incremental M-step
  - Variational EM: relax E-step from maximizing to increasing $\mathcal{L}$
    - no longer $\mathcal{L} = \ln p(\mathbf{X}|\theta)$ after E-step

- Same applies for MAP estimation $p(\theta|\mathbf{X}) = p(\theta)p(\mathbf{X}|\theta)/p(\mathbf{X})$

  - bound second term: $\ln p(\theta|\mathbf{X}) \geq \ln p(\theta) + \mathcal{L}(q, \theta) - \ln p(\mathbf{X})$