

COMS30035, Machine learning: The EM algorithm

James Cussens

School of Computer Science
University of Bristol

11th October 2024

MLE for a Gaussian mixture

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

- ▶ No closed form for the MLE
- ▶ (At least $K!$ solutions)
- ▶ So have to resort to an iterative algorithm.
- ▶ “It should be emphasized that there will generally be multiple local maxima of the log likelihood function” [Bis06, p. 438].
- ▶ The iterative algorithm we use is called the *Expectation-Maximization (EM) algorithm*.

Settings derivatives to zero

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

$$\pi_k = \frac{N_k}{N}$$

where $\gamma(z_{nk}) = p(z_k = 1 | \mathbf{x}_n)$ and $N_k = \sum_{n=1}^N \gamma(z_{nk})$.

- See [Bis06, §9.22] for the derivation.

EM for Gaussian mixtures

- To initialise the EM algorithm we choose starting values for μ , Σ and π .

E step Compute the values for the responsibilities $\gamma(z_{nk})$ given the current parameter values:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)}$$

M step Re-estimate the parameters using the current responsibilities:

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^T$$

$$\pi_k^{\text{new}} = \frac{N_k}{N}$$

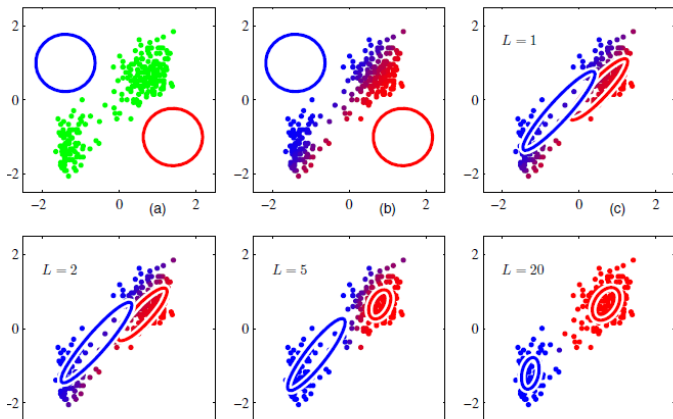
This E-step is just Bayes theorem

$$p(z_k = 1 | \mathbf{x}_n) = \frac{p(z_k = 1)p(\mathbf{x}_n | z_k = 1)}{p(\mathbf{x}_n)} = \frac{p(z_k = 1)p(\mathbf{x}_n | z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(\mathbf{x}_n | z_j = 1)}$$

The same equation in different notation is:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

EM in pictures



Why does EM work?

- ▶ We have yet to show that each iteration of the EM algorithm increases the log-likelihood $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- ▶ We will do this for the general case:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{z}} p(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta}) \right\}$$

- ▶ Z are hidden variables (i.e. not observed) also called *latent variables*.
- ▶ $\{X, Z\}$ is the *complete data*. Assume that if we had the complete data then MLE would be easy.
- ▶ $\{X\}$ is the *incomplete data*.

Decomposing the log-likelihood

- ▶ Let $q(\mathbf{Z})$ be any distribution over the hidden variables.
- ▶ We have the following key decomposition of the log-likelihood:

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \text{KL}(q||p)$$

where

$$\begin{aligned}\mathcal{L}(q, \boldsymbol{\theta}) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \\ \text{KL}(q||p) &= - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}\end{aligned}$$

- ▶ An exercise for you: prove that this decomposition is correct (Exercise 9.24 in Bishop). Use the tip Bishop gives on p.451.

Kullback-Leibler divergence

$$\text{KL}(q||p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \right\}$$

- ▶ $\text{KL}(p_1||p_2)$ denotes the *Kullback-Leibler divergence* between probability distributions p_1 and p_2 .
- ▶ KL-divergence is important in, e.g., information theory.
- ▶ It's a bit like a 'distance' between two distributions.
- ▶ But it is not a true distance since, for example, it is not symmetric.
- ▶ $\text{KL}(p_1||p_2) \geq 0$ and $\text{KL}(p_1||p_2) = 0$ if and only if $p_1 = p_2$.

EM: key ideas

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \text{KL}(q||p)$$

- ▶ $\text{KL}(q||p) \geq 0$ for any choice of q , so $\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$.
- ▶ In the E-step we increase $\mathcal{L}(q, \boldsymbol{\theta})$ by updating q (and leaving $\boldsymbol{\theta}$ fixed).
- ▶ In the M-step we increase $\mathcal{L}(q, \boldsymbol{\theta})$ by updating $\boldsymbol{\theta}$ (and leaving q fixed).
- ▶ After the E-step we have $\mathcal{L}(q, \boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta})$, so that in the following M-step increasing $\mathcal{L}(q, \boldsymbol{\theta})$ will also increase $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

The E-step

$$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}}) = \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \text{KL}(q||p)$$

- ▶ In the E-step we update q but leave $\boldsymbol{\theta}^{\text{old}}$ fixed.
- ▶ $\text{KL}(q||p) = 0$ when $q = p$, so to maximise $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ we set $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
- ▶ This increases $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ but not $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$.
- ▶ [Bis06, Fig 9.12] illustrates the E-step.

The M-step

$$\ln p(\mathbf{X}|\theta^{\text{new}}) = \mathcal{L}(q, \theta^{\text{new}}) + \text{KL}(q||p)$$

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

- ▶ In the M-step we find parameters θ^{new} which maximise $\mathcal{L}(q, \theta)$, while leaving q fixed.
- ▶ This will necessarily increase $\ln p(\mathbf{X}|\theta)$ since $\text{KL}(q||p) \geq 0$.
- ▶ In fact we get a 'bonus' since changing p from $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ to $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{new}})$ will (typically) lead $\text{KL}(q||p)$ to increase from 0 to some positive value.
- ▶ [Bis06, Fig 9.13] illustrates the M-step.

Visualising EM

- ▶ David Barber (from UCL) provides a nice visualisation for EM (where the latent variable is binary).
- ▶ It's Fig 11.2 on p.259 of his [book](#) [Bar12].

Back to Gaussian mixtures

- ▶ In the standard case of independent and identically distributed (i.i.d.) dataset \mathbf{X} , we get:

$$p(\mathbf{Z}|\mathbf{X}, \theta) = \prod_{n=1}^N p(\mathbf{z}_n|\mathbf{x}_n, \theta)$$

- ▶ In the case of Gaussian mixtures the responsibilities $\gamma(z_{nk})$ define the $p(\mathbf{z}_n|\mathbf{x}_n, \theta)$.
- ▶ So computing the responsibilities is the E-step.
- ▶ And the M-step we saw on slide 4 does indeed maximise $\mathcal{L}(q, \theta)$ given the current responsibilities.
- ▶ Proving this is Exercises 9.8 and 9.9 in Bishop.

Reading

- ▶ Bishop §9.2.2
- ▶ Bishop §9.4
- ▶ Murphy §8.7.2-§8.7.3

Problems and quizzes

- ▶ Bishop Exercise 9.24 (Use the tip Bishop gives on p. 451)
- ▶ Quizzes:
 - ▶ Week 5: The EM algorithm



David Barber.

Bayesian Reasoning and Machine Learning.

Cambridge University Press, 2012.



Christopher M. Bishop.

Pattern Recognition and Machine Learning.

Springer, 2006.