Machine Learning

Expectation Maximisation

Kia Nazarpour

GMM: Reminder

- Hard boundaries are exchanged for flexible and probabilistic soft boundaries
- Immense flexibility: $p(\mathbf{x}_n|\cdots)$ can take the form of any probability density including Bernoulli distribution
- Expectation Maximisation instead of Maximum Likelihood

Learning Outcomes

1. Move from Gaussian Mixture Models to Latent Variable Models (abstraction)
2. Understand the key motivation behind Expectation Maximisation (EM).
3. Review observed and latent variables.
4. Study the EM formula

References:

1. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2008. (Section 9.4)

General latent Variable model

- Two sets of random variables $\mathbf{X}$ and $\mathbf{Z}$.
- $\mathbf{X}$ captures all observed variables.
- $\mathbf{Z}$ captures all unseen/hidden/latent/unobserved variables
- Joint probability model is parametrised by $\boldsymbol{\theta} \in \Theta$ as

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

EM - Key motivation

- It is hard to optimise for marginal log-likelihood
- Typically, it is easier to optimise the log-likelihood for the complete data

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{X}|\boldsymbol{\theta})$$

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Jensen's Inequality - Reminder

Theorem

If $f : \mathcal{R} \to \mathcal{R}$ is a convex function and x is a random variable, then

$$\mathbb{E}\, f(x) \geq f\, \mathbb{E}\, x$$

- For example $f(x) = x^2$ is a convex function and $\mathbb{E}\, x^2 \geq (\mathbb{E}\, x)^2$

$$\sigma^2(x) = \mathbb{E}\, x^2 - (\mathbb{E}\, x)^2 \geq 0$$

Kullback-Leibler Divergence - Reminder

- For discrete probability distributions p and q on the same probability space $\mathcal{X}$
- The KL-divergence is defined by

$$\mathrm{KL}(p\|q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right] = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

- The KL-divergence measures the "distance" between p and q but
- The KL-divergence **is not a metric**.
- The KL-divergence **is not a symmetric**.

$$\begin{aligned} \mathrm{KL}(p\|q) &\geq 0 \\ \mathrm{KL}(p\|q) &\neq \mathrm{KL}(q\|p) \\ \mathrm{KL}(p\|p) &= 0 \end{aligned}$$

EM - Key motivation

- It is hard to optimise for marginal log-likelihood
- Typically, it is easier to optimise the log-likelihood for the complete data

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{X}|\boldsymbol{\theta})$$

$$\max_{\boldsymbol{\theta}} \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Lower Bound for Marginal Log-likelihood

Let $q(\mathbf{X})$ be any discrete probability function on $\mathcal{Z}$

$$\log p(\mathbf{X}|\boldsymbol{\theta}) \quad = \quad \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Lower Bound for Marginal Log-likelihood

Let $q(\mathbf{Z})$ be any discrete probability function on $\mathcal{Z}$

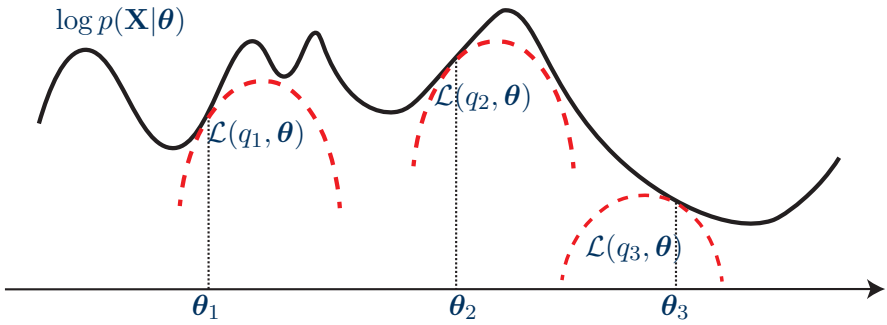
$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\theta}) &= \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\ &= \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right)\end{aligned}$$

Lower Bound for Marginal Log-likelihood

Let $q(\mathbf{Z})$ be any discrete probability function on $\mathcal{Z}$

$$\begin{aligned}
\log p(\mathbf{X}|\boldsymbol{\theta}) &= \log \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\
&= \log \sum_{\mathbf{Z}} q(\mathbf{Z}) \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right) \\
&\geq \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right)}_{\mathcal{L}(q, \boldsymbol{\theta})}
\end{aligned}$$

EM - Visualisation - 1



Lower Bound for Marginal Log-likelihood

Let $q(\mathbf{Z})$ be any discrete probability function on $\mathcal{Z}$

$$\log p(\mathbf{X}|\boldsymbol{\theta}) \geq \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right)}_{\mathcal{L}(q, \boldsymbol{\theta})}$$

$$\text{evidence} \geq \textbf{E}\text{vidence } \textbf{L}\text{ower } \textbf{BO}\text{und}$$

In EM, we maximise the ELBO w.r.t. to q and $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}}_{\text{EM}} = \arg\max_{\boldsymbol{\theta}} \left(\arg\max_{q} \mathcal{L}(q, \boldsymbol{\theta}) \right)$$

EM - Visualisation - 2

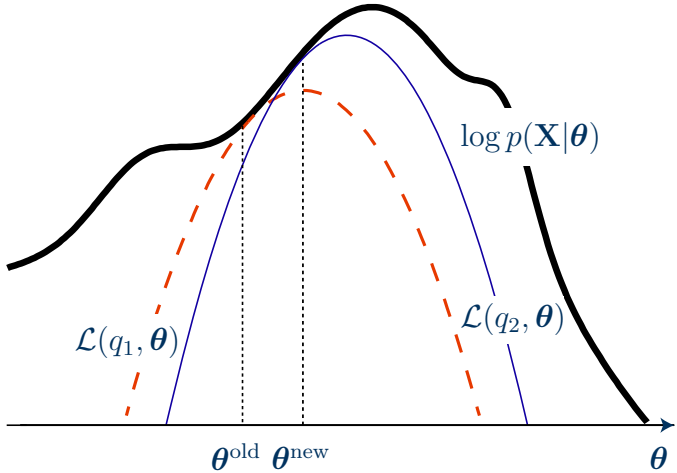


Figure based on Figure 9.14 of Bishop (2008).

ELBO reformulation

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

$$
\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}[q \| p]
$$

EM - Visualisation

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}[q\|p]$$

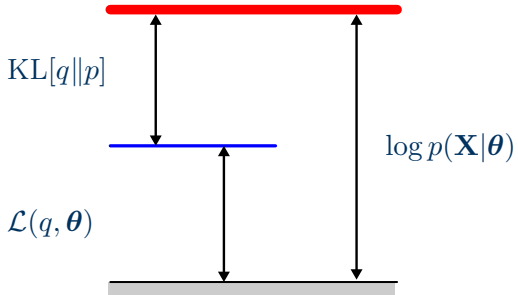


Figure 9.11 of Bishop (2008).

EM - Visualisation

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}[q\|p]$$

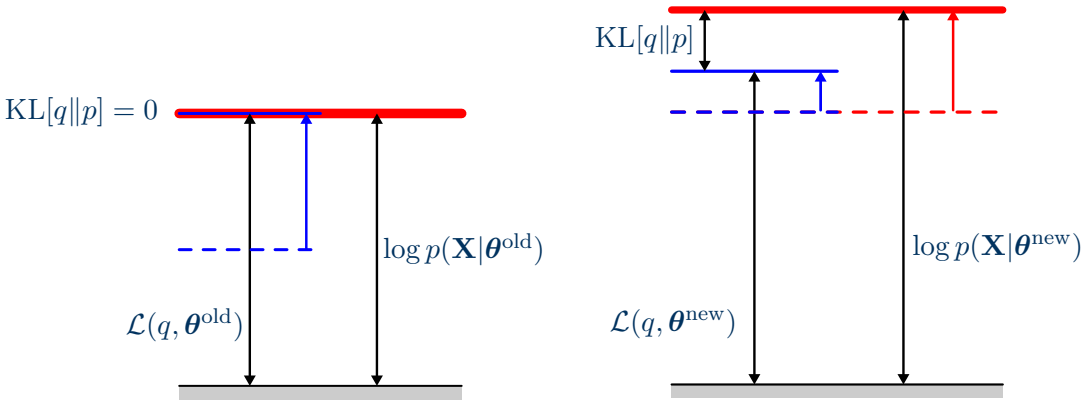


Figure 9.12-3 of Bishop (2008).

EM - Summary - 1

1. Choose an initial $\boldsymbol{\theta}^{\text{old}}$
2. **Expectation Step**
   - Let $q^*(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$, giving the best lower bound at $\boldsymbol{\theta}^{\text{old}}$
   - Let

$$J(\boldsymbol{\theta}) := (q^*, \boldsymbol{\theta}) = \underbrace{\sum_{\mathbf{Z}} q^*(\mathbf{Z}) \log \left(\frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q^*(\mathbf{Z})} \right)}_{\text{Expectation}}$$

3. **Maximisation Step**

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$$

4. Go to step 2 until convergence

EM - Summary - 2

1. Maximum likelihood estimation is easy if we observe all the values of all the relevant random variables.
2. In case of missing data and/or latent variables, then Maximum likelihood estimation becomes hard.
3. In such cases, it is often simpler (but not always faster) to use the EM algorithm.
4. EM alternates between inferring the missing values given the parameters (E step), and then optimising the parameters given the *filled* in data (M step).
5. EM monotonically increases the observed data log likelihood.