Introduction to Statistical Machine Learning

Cheng Soon Ong & Christian Walder

Machine Learning Research Group
Data61 | CSIRO
and
College of Engineering and Computer Science
The Australian National University

Canberra
February – June 2019

(Many figures from C. M. Bishop, "Pattern Recognition and Machine Learning")

Part VIII

Mixture Models and EM 2

Recall: K-means Clustering

- Find the values for $\{r_{nk}\}$ and $\{\boldsymbol{\mu}_k\}$ so as to minimise J.
- But $\{r_{nk}\}$ depends on $\{\boldsymbol{\mu}_k\}$, and $\{\boldsymbol{\mu}_k\}$ depends on $\{r_{nk}\}$.
- Iterate until no further change
  1. Minimise J w.r.t. r_{nk} while keeping $\{\boldsymbol{\mu}_k\}$ fixed,

$$r_{nk} = \begin{cases} 1, & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0, & \text{otherwise.} \end{cases} \qquad \forall n = 1, \dots, N$$

     Expectation step

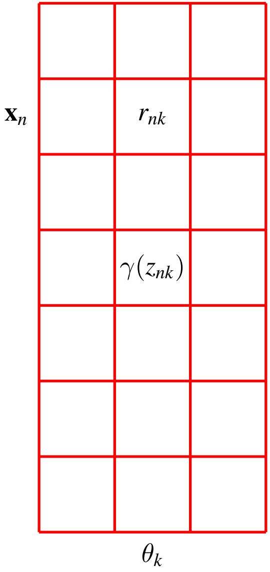
  2. Minimise J w.r.t. $\{\boldsymbol{\mu}_k\}$ while keeping r_{nk} fixed,

$$0 = 2 \sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} r_{nk}\mathbf{x}_n}{\sum_{n=1}^{N} r_{nk}}$$

     Maximisation step

Responsibilities



For k-means clustering,

we have hard assignments

For GMM,

we have soft assignments

Ong & Walder & Webers
Data61 | CSIRO
The Australian National University

EM for Gaussian Mixtures

- Starting point is the log of the likelihood function

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

- Critical point of $\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ w.r.t. $\boldsymbol{\mu}_k$

$$0 = \sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

- Therefore

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

where the effective number of points assigned to Gaussian k is $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$.

EM for Gaussian Mixtures

- Maximum of the log of the likelihood function for

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

- Similarly for the covariance matrix

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

- and for the mixing coefficients π_k (using a Lagrange multiplier as $\sum_k \pi_k = 1$)

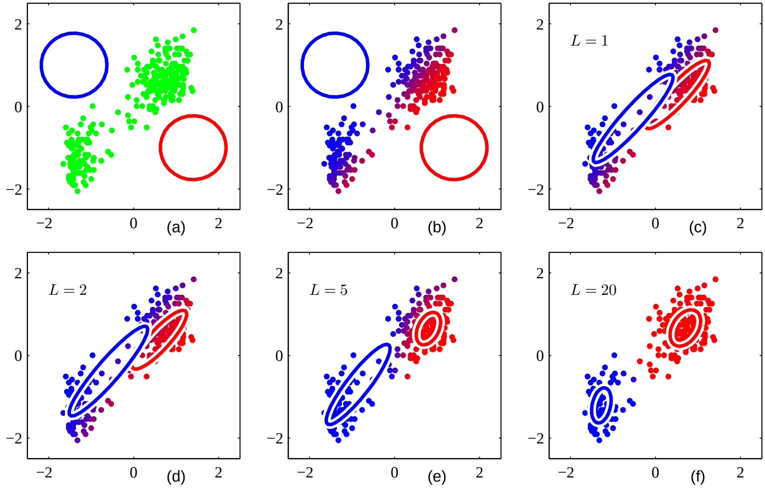
$$\pi_k = \frac{N_k}{N}.$$

- This is not a closed form solution because the responsibilities $\gamma(z_{nk})$ depend on $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$.

EM for Gaussian Mixtures

- Given a Gaussian mixture and data $\mathbf{X}$, maximise the log likelihood w.r.t. the parameters $(\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$.
  1. Initialise the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k. Evaluate the log likelihood function.
  2. E step : Evaluate the $\gamma(z_{nk})$ using the current parameters
  $$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$
  3. M step : Re-estimate the parameters using the current $\gamma(z_{nk})$
  $$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})\, \mathbf{x}_n \qquad \pi_k^{\text{new}} = \frac{N_k}{N}$$
  $$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})(\mathbf{x}_n - \boldsymbol{\mu}_k^{\text{new}})^T$$
  4. Evaluate the log likelihood, if not converged then goto 2.
  $$\ln p(\mathbf{X} \mid \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k^{\text{new}} \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}) \right\}$$

EM for Gaussian Mixtures - Example



EM for Gaussian Mixtures - Relation to K-Means

- Assume a Gaussian mixture model.
- Covariance matrices given by $\epsilon\mathbf{I}$, where ϵ is shared by all components.
- Then

$$p(\mathbf{x}\,|\,\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}-\boldsymbol{\mu}_k\|^2\right\}.$$

- Keep ϵ fixed, do not re-estimate.
- Responsibilities

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n-\boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n-\boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$

- Taking the limit $\epsilon \to 0$, the term in the denominator for which $\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2$ is the smallest will go to zero most slowly.

EM for Gaussian Mixtures - Relation to K-Means

- Assume a Gaussian mixture model.

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}$$

- Therefore

$$\gamma(z_{nk}) = \begin{cases} 1 & \text{if } \|\mathbf{x}_n - \boldsymbol{\mu}_k\| < \|\mathbf{x}_n - \boldsymbol{\mu}_j\| \qquad \forall j \neq k \\ 0 & \text{otherwise} \end{cases}$$

- Holds independent of π_k as long as none are zero.
- Hard assignment to exactly one cluster : K-means.

$$\lim_{\epsilon \to 0} \gamma(z_{nk}) = r_{nk}$$

Mixture of Bernoulli Distributions

- Set of D binary variables x_i, $i = 1, \dots, D$.
- Each governed by a Bernoulli distribution with parameter μ_i. Therefore

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

- Expectation and covariance

$$\mathbb{E}\left[\mathbf{x}\right] = \boldsymbol{\mu}$$
$$\operatorname{cov}[\mathbf{x}] = \operatorname{diag}\{\mu_i(1-\mu_i)\}$$

Mixture of Bernoulli Distributions

- Mixture

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \pi_k \, p(\mathbf{x} \mid \boldsymbol{\mu}_k)$$

with

$$p(\mathbf{x} \mid \boldsymbol{\mu}_k) = \prod_{i=1}^{D} \mu_{ki}^{x_i} (1 - \mu_{ki})^{1-x_i}$$

- Similar calculation as with mixture of Gaussian

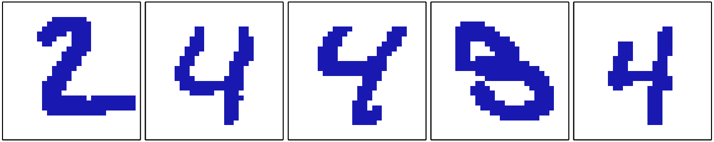
$$\gamma(z_{nk}) = \frac{\pi_k \, p(\mathbf{x}_n \mid \boldsymbol{\mu}_k)}{\sum_{j=1}^{K} \pi_j \, p(\mathbf{x}_n \mid \boldsymbol{\mu}_j)}$$

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

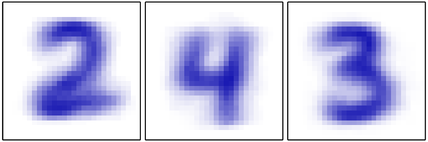
$$\bar{\mathbf{x}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \qquad \mu_k = \bar{\mathbf{x}}$$

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

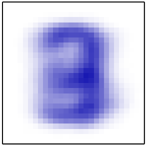
EM for Mixture of Bernoulli Distributions - Digits



Examples from a digits data set, each pixel taken only binary values.



Parameters μ_{ki} for each component in the mixture.



Fit to one multivariate Bernoulli distribution.

The Role of Latent Variables

- EM finds the maximum likelihod solution for models with latent variables.
- Two kinds of variables
  - Observed variables $\mathbf{X}$
  - Latent variables $\mathbf{Z}$

  plus model parameters $\boldsymbol{\theta}$.
- Log likelihood is then

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

- Optimisation problem due to the log-sum.
- Assume maximisation of the distribution $p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$ over the complete data set $\{\mathbf{X}, \mathbf{Z}\}$ is straightforward.
- But we only have the incomplete data set $\{\mathbf{X}\}$ and the posterior distribution $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta})$.

EM - Key Idea

- Key idea of EM: As $\mathbf{Z}$ is not observed, work with an 'averaged' version $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ of the complete log-likelihood $\ln p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$, averaged over all states of $\mathbf{Z}$.

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$$

EM Algorithm

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. E step Evaluate $p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. M step Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \, \ln p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})$$

4. Check for convergence of log likelihood or parameter values. If not yet converged, then

$$\boldsymbol{\theta}^{\text{old}} = \boldsymbol{\theta}^{\text{new}}$$

   and go to step 2.

EM Algorithm - Convergence

- Start with the product rule for the observed variables $\mathbf{X}$, the unobserved variables $\mathbf{Z}$, and the parameters $\boldsymbol{\theta}$

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

- Apply $\sum_{\mathbf{Z}} q(\mathbf{Z})$ with arbitrary $q(\mathbf{Z})$ to the formula

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta}) + \ln p(\mathbf{X} \,|\, \boldsymbol{\theta}).$$

- Rewrite as

$$\ln p(\mathbf{X} \,|\, \boldsymbol{\theta}) = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \,|\, \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathcal{L}(q, \boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} \,|\, \mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathrm{KL}(q \| p)}$$

- $\mathrm{KL}(q \| p)$ is the Kullback-Leibler divergence.

Kullback-Leibler Divergence

- 'Distance' between two distributions $p(y)$ and $q(y)$

$$\mathrm{KL}(q\|p) = \sum_y q(y) \ln \frac{q(y)}{p(y)} \qquad = -\sum_y q(y) \ln \frac{p(y)}{q(y)}$$

$$\mathrm{KL}(q\|p) = \int q(y) \ln \frac{q(y)}{p(y)} \,\mathrm{d}y \qquad = -\int q(y) \ln \frac{p(y)}{q(y)} \,\mathrm{d}y$$

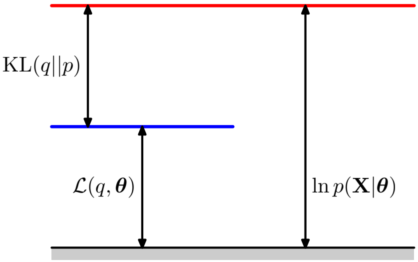
- $\mathrm{KL}(q\|p) \geq 0$
- not symmetric: $\mathrm{KL}(q\|p) \neq \mathrm{KL}(p\|q)$
- $\mathrm{KL}(q\|p) = 0$ iff $q = p$.
- invariant under parameter transformations
- Example: Kullback-Leibler divergence between two normal distributions $q(x) = \mathcal{N}(x \mid \mu_1, \sigma_1)$ and $p(x) = \mathcal{N}(x \mid \mu_2, \sigma_2)$

$$\mathrm{KL}(q\|p) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$$

EM Algorithm - Convergence

- The two parts of $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$

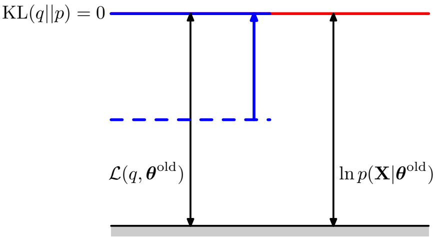
$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathcal{L}(q, \boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathrm{KL}(q \| p)}$$



EM Algorithm - E Step

- Hold $\boldsymbol{\theta}^{\text{old}}$ fixed. Maximise the lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ with respect to $q(\cdot)$.
- $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ is a functional.
- $\ln p(\mathbf{X} \mid \boldsymbol{\theta})$ does NOT depend on $q(\cdot)$.
- Maximum for $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ will occur when the Kullback-Leibler divergence vanishes.
- Therefore, choose $q(\mathbf{Z}) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$

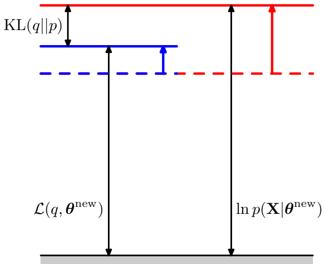
$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathcal{L}(q,\boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathrm{KL}(q\|p)}$$



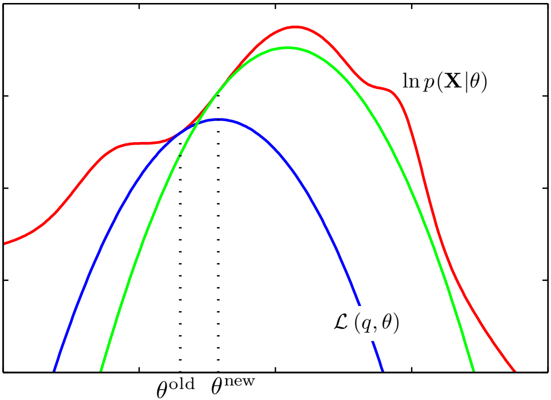
EM Algorithm - M Step

- Hold $q(\cdot) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ fixed. Maximise the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$:
  $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\theta} \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) = \arg\max_{\theta} \sum_{\mathbf{Z}} q(\cdot) \ln p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})$
- $\mathcal{L}(q, \boldsymbol{\theta}^{\text{new}}) > \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ unless maximum already reached.
- As $q(\cdot) = p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ is fixed, $p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta}^{\text{new}})$ will not be equal to $q(\cdot)$, and therefore the Kullback-Leiber distance will be greater than zero (unless converged).

$$\ln p(\mathbf{X} \mid \boldsymbol{\theta}) = \underbrace{\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathcal{L}(q, \boldsymbol{\theta})} \underbrace{- \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})}}_{\mathrm{KL}(q \| p)}$$



EM Algorithm - Parameter View



Red curve : incomplete data likelihood.
Blue curve : After E step. Green curve : After M step.