annotated version

Machine Learning Course - CS-433

Expectation-Maximization Algorithm

Nov 16, 2017

minor changes by Martin Jaggi 2016
minor changes by Martin Jaggi 2017

Last updated: November 16, 2017

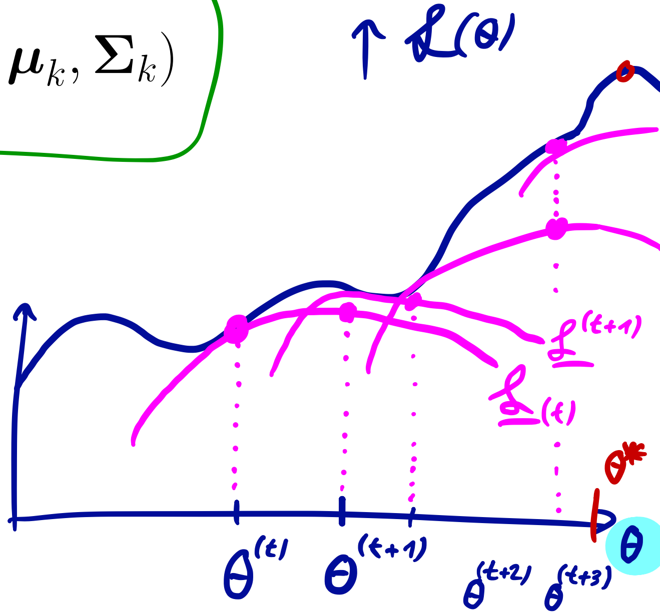
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

Computing maximum likelihood for Gaussian mixture model is difficult due to the log outside the sum.

$$\max_{\boldsymbol{\theta}} \ \mathcal{L}(\boldsymbol{\theta}) := \sum_{n=1}^{N} \overbrace{\log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}^{\mathcal{L}_n}$$

Expectation-Maximization (EM) algorithm provides an elegant and general method to optimize such optimization problems. It uses an iterative two-step procedure where individual steps usually involve problems that are easy to optimize.



EM algorithm: Summary

Start with $\boldsymbol{\theta}^{(1)}$ and iterate:

1. Expectation step: Compute a lower bound to the cost such that it is tight at the previous $\boldsymbol{\theta}^{(t)}$:

model of $\mathcal{L}(\theta)$

$$\mathcal{L}(\boldsymbol{\theta}) \geq \underline{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}) \text{ and}$$
$$\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \underline{\mathcal{L}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)}).$$

- lower bound to $\mathcal{L}$ for all θ
- coincides with $\mathcal{L}$ at $\theta = \theta^{(t)}$

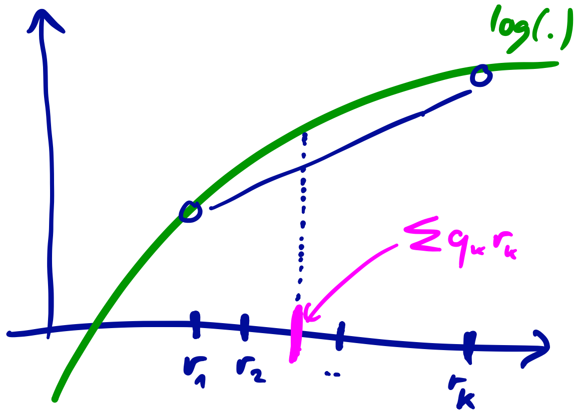
2. Maximization step: Update $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} \underline{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)}).$$

How to define $\boxed{\mathcal{L}(\theta,\theta^{(t)})}$?

Concavity of log

Given non-negative weights q s.t. $\sum_k q_k = 1$, the following holds for any $r_k > 0$:



$$\log\left(\sum_{k=1}^{K} q_k r_k\right) \geq \sum_{k=1}^{K} q_k \log r_k$$

Jensen's Inequality
$\Leftarrow$ log is concave

① The expectation step

$$\overbrace{\log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}^{\mathcal{L}_n(\theta)} \geq \overbrace{\sum_{k=1}^{K} q_{kn}^{(t)} \log \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{q_{kn}^{(t)}}}_{r_k}}^{=: \mathcal{L}(\theta,\theta^{(t)})}$$

- lower bound: from $\geq$
- coincides with $\mathcal{L}$ at $\theta^{(t)}$?

with equality when,

$$q_{kn}^{(t)} = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}$$

This is not a coincidence.

$$\mathcal{L}(\theta^{(t)},\theta^{(t)}) = \sum_{k=1}^{K} \underbrace{\frac{\pi_k \mathcal{N}}{\sum_{k'} \pi_{k'} \mathcal{N}}}_{q_{kn}} \log \frac{\cancel{\pi_k \mathcal{N}}}{\underbrace{\frac{\cancel{\pi_k \mathcal{N}}}{\sum_{k'} \pi_{k'} \mathcal{N}}}_{q_{kn}}}$$

$$= \log \sum_k \pi_k \mathcal{N}$$

$$= \mathcal{L}(\theta^{(t)})$$

②

The maximization step

Maximize the lower bound w.r.t. $\boldsymbol{\theta}$.

$$\max_{\boldsymbol{\theta}} \underbrace{\sum_{n=1}^{N}\sum_{k=1}^{K} q_{kn}^{(t)}\left[\log \pi_k + \log \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right] - \cancel{\log q_{kn}^{(t)}}}_{\mathcal{L}(\theta, \theta^{(t)})}$$

$\sum_k \left[\sum_n q \left[\quad \right] \right]$

independent of θ

Differentiating w.r.t. $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k^{\cancel{-1}}$, we can get the updates for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$.

$$\nabla_{\mu_k} \mathcal{L}(\theta, \theta^{(t)}) \overset{!}{=} 0$$

$$\nabla_{\Sigma_k} \mathcal{L}(\theta, \theta^{(t)}) \overset{!}{=} 0$$

$$\boldsymbol{\mu}_k^{(t+1)} := \frac{\sum_n q_{kn}^{(t)} \mathbf{x}_n}{\sum_n q_{kn}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} := \frac{\sum_n q_{kn}^{(t)} \boxed{(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^\top}}{\sum_n q_{kn}^{(t)}}$$

$= \boxed{v} \times \boxed{v^\top}$

For π_k, we use the fact that they sum to 1. Therefore, we add a Lagrangian term, differentiate w.r.t. π_k and set to 0, to get the following update:

want $\sum \pi_k = 1$

$$\underbrace{\ldots + \beta\left(\sum \pi_k - 1\right)}$$

$\hookrightarrow \nabla_{\pi} \ldots = 0$

$$\pi_k^{(t+1)} := \frac{1}{N}\sum_{n=1}^{N} q_{kn}^{(t)}$$

Summary of EM for GMM

Initialize $\boldsymbol{\mu}^{(1)}, \boldsymbol{\Sigma}^{(1)}, \boldsymbol{\pi}^{(1)}$ and iterate between the E and M step, until $\mathcal{L}(\boldsymbol{\theta})$ stabilizes.

In the limit when $\Sigma_k = \sigma I$ and $\sigma \to 0$:

1. E-step: Compute assignments $q_{kn}^{(t)}$:

$\approx \; .. \exp(..-\|x_n - \mu_k\|^2)$

$$q_{kn}^{(t)} := \frac{\pi_k^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{k=1}^{K} \pi_k^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}$$

largest for which k?

k-means $\to$ assignment z_{kn} (closest center)

2. Compute the marginal likelihood (cost).

$$\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k^{(t)} \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})$$

3. M-step: Update $\boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}, \pi_k^{(t+1)}$.

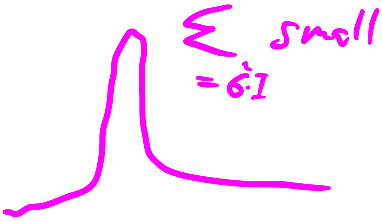
$$\boldsymbol{\mu}_k^{(t+1)} := \frac{\sum_n q_{kn}^{(t)} \mathbf{x}_n}{\sum_n q_{kn}^{(t)}}$$

k-means: cluster means

$$\boldsymbol{\Sigma}_k^{(t+1)} := \frac{\sum_n q_{kn}^{(t)} (\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_n - \boldsymbol{\mu}_k^{(t+1)})^\top}{\sum_n q_{kn}^{(t)}}$$

$$\pi_k^{(t+1)} := \frac{1}{N} \sum_n q_{kn}^{(t)}$$

If we let the covariance be diagonal i.e. $\boldsymbol{\Sigma}_k := \sigma^2 \mathbf{I}$, then EM algorithm is same as K-means as $\sigma^2 \to 0$.



E-step: compute q_{kn}

M-step: update parameters μ, Σ, π

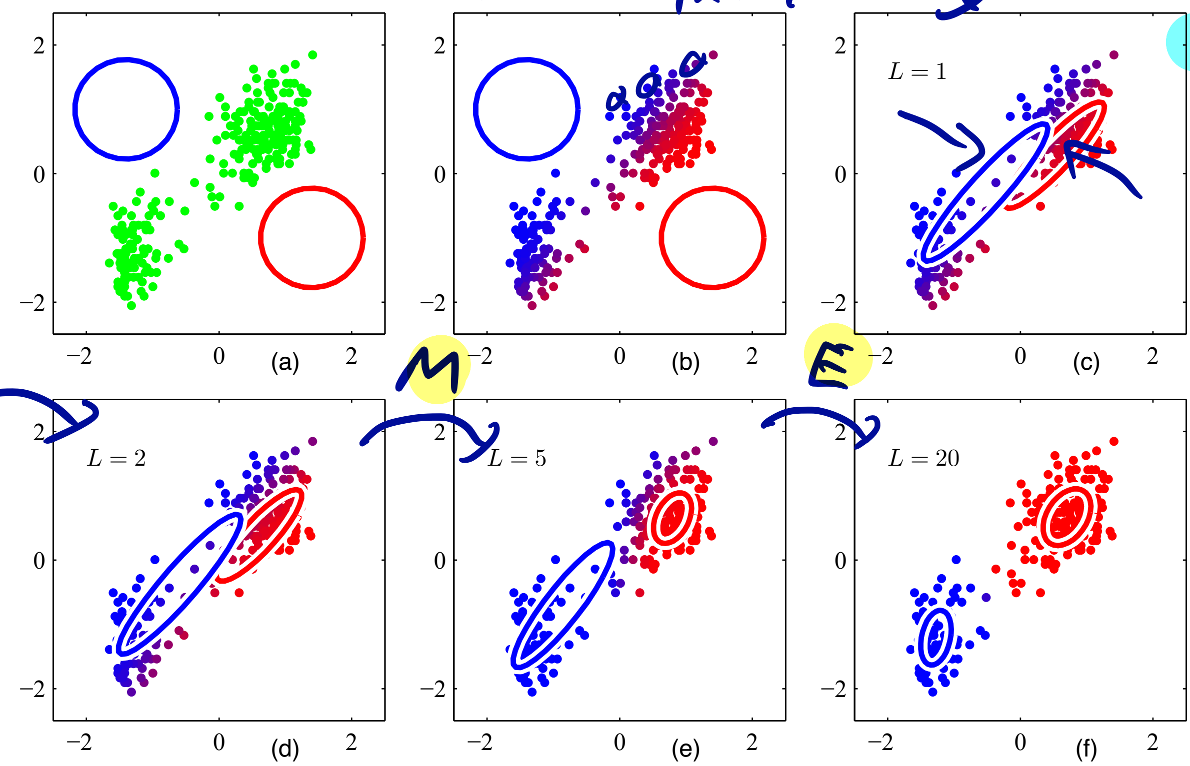


Figure 1: EM algorithm for GMM

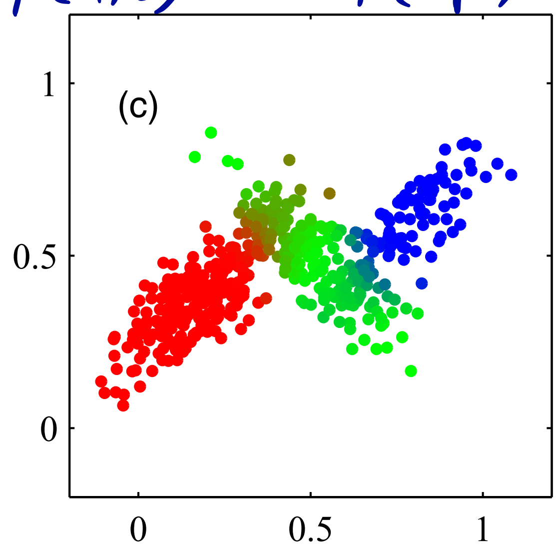
Posterior distribution

We now show that $q_{kn}^{(t)}$ is the posterior distribution of the latent variable, i.e. $q_{kn}^{(t)} = p(z_n = k \,|\, \mathbf{x}_n, \boldsymbol{\theta}^{(t)})$ posterior

joint = Likelihood · prior = posterior · marginal likelihood

$$p(\mathbf{x}_n, z_n|\boldsymbol{\theta}) = p(\mathbf{x}_n|z_n, \boldsymbol{\theta})p(z_n|\boldsymbol{\theta}) = p(z_n|\mathbf{x}_n, \boldsymbol{\theta})p(\mathbf{x}_n|\boldsymbol{\theta})$$

$$p(A,B) = p(A|B) \cdot p(B) = p(B|A) \cdot p(A)$$



$$p(z_n = k|x_n, \theta) = \frac{LH \cdot prior}{ML}$$

$$= \frac{p(z_n = k|\theta) \cdot p(x_n|z_n = k, \theta)}{\sum_j p(z_n = j|\theta) \cdot p(x_n|z_n = j, \theta)} \quad = p(x_n)$$

$$= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

$$=: q_{kn}$$

EM in general

Given a general joint distribution $p(\mathbf{x}_n, z_n|\boldsymbol{\theta})$, the marginal likelihood can be lower bounded similarly:

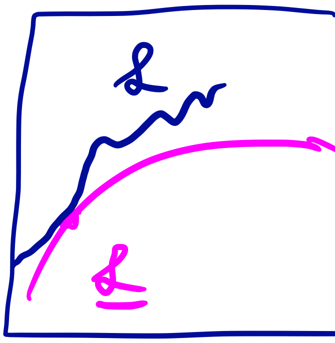
$\max_\theta \; p(x_n|\theta)$

$\sum_n \log(\mathbb{E}_z \ldots) \rightarrow \sum \mathbb{E}_z \log(\ldots)$

The EM algorithm can be compactly written as follows:

$$\boldsymbol{\theta}^{(t+1)} := \arg\max_{\boldsymbol{\theta}} \sum_{n=1}^{N} \mathbb{E}_{p(z_n|\mathbf{x}_n,\boldsymbol{\theta}^{(t)})}\big[\log p(\mathbf{x}_n, z_n|\boldsymbol{\theta})\big]$$

Example: GMM: $= \sum_n \sum_k \underbrace{p(z_n=k|x_n,\theta^{(t)})}_{\text{posterior} \overset{\text{GMM}}{=} q_{kn}} \log p(x_n, z_n|\theta)$



→ EM as before

Another interpretation is that part of the data is missing, i.e. $(\mathbf{x}_n, z_n)$ is the "complete" data and z_n is missing. The EM algorithm averages over the "unobserved" part of the data.

ToDo

1. Identify the joint, likelihood, prior, and marginal distributions respectively. Understand the use of Bayes rule that relates all these distributions together.
2. Derive the posterior distribution for GMM.
3. Understand the relation between EM and K-means.
4. Relate the lower bound to EM for probabilistic models in general.
5. Read the Wikipedia page on how to find a good K.
6. Read about other mixture models in the KPM book.