annotated version

Machine Learning Course - CS-433

Expectation-Maximization Algorithm

find θ

Nov 17, 2016

minor changes by Martin Jaggi 2016

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

$\theta = (\pi_{1:K}, \mu_{1:K}, \Sigma_{1:K})$

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

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

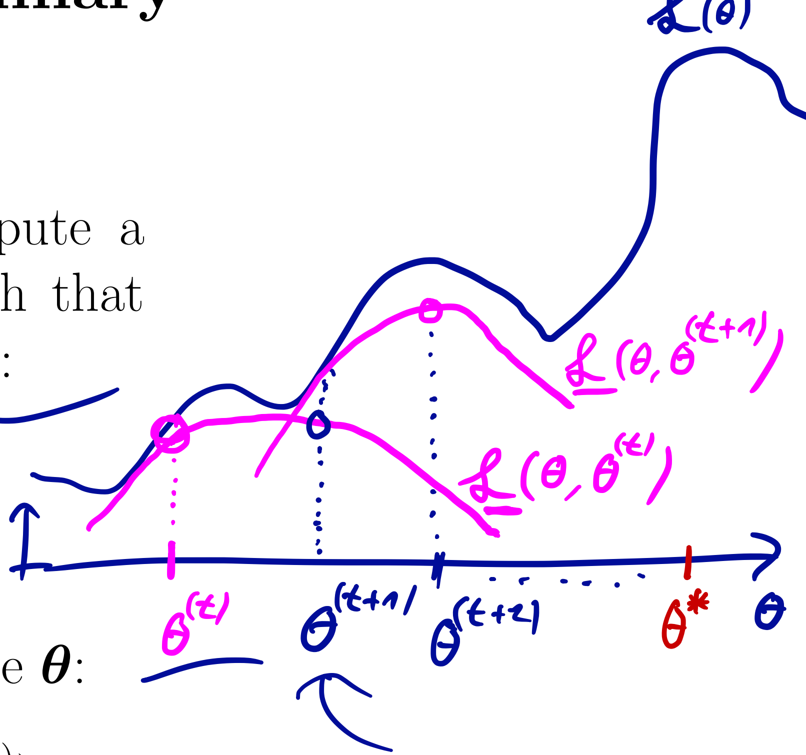
Expectation-Maximization (EM) algorithm provides an elegant and general method to optimize such optimization problems. It uses an iterative two-step procedure where individual steps usually involve problems that are easy to optimize.

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 $\boldsymbol{\theta}^{(t)}$:

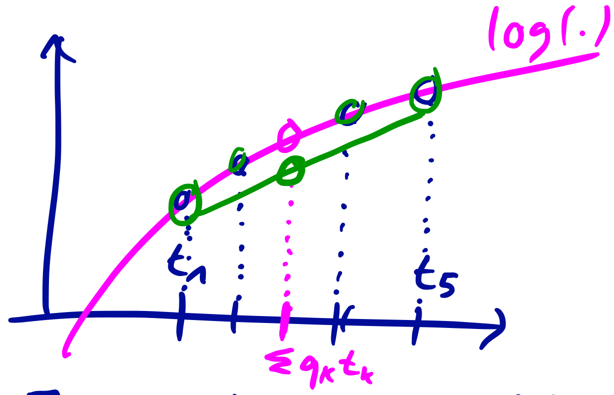
$\mathcal{L}(\boldsymbol{\theta}) \geq \boxed{\underline{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})}$ and
$\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \underline{\mathcal{L}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})$.



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

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

Concavity of log



Jensen's Inequality

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

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

The expectation step

(per data example)

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

with equality when,

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

This is not a coincidence.

$$= \log \sum_k \underbrace{\boxed{\frac{\pi_k \mathcal{N}(x_n | \mu, \Sigma)}{q_{kn}}}}_{= t_k} q_{kn}$$

$$\underline{\mathcal{L}}(\theta^{(t)}, \theta^{(t)}) \stackrel{?}{=} \mathcal{L}(\theta^{(t)})$$

$$= \sum_k q_{kn}^{(t)} \log \frac{\pi^{(t)} \mathcal{N}(x, \theta^{(t)})}{q_{kn}^{(t)}}$$

$\downarrow$

$$= \sum_k \frac{\pi \mathcal{N}^{(k)}}{\sum \pi \mathcal{N}} \log \frac{\cancel{\pi \mathcal{N}}^{(k)}}{\frac{\cancel{\pi \mathcal{N}}^{(k)}}{\sum \pi \mathcal{N}}}$$

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

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

for fixed $q_{kn}^{(t)}$

The maximization step

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

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

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

$$\boxed{q \log \frac{\pi \mathcal{N}(..)}{q}}$$

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

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

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

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

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

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:

$$\ldots + \beta \left(\sum_k \pi_k - 1 \right)$$

$$\hookrightarrow \nabla_{..} \stackrel{!}{=} 0$$

$$\hookrightarrow \pi$$

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

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.

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

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

for $\Sigma_k = \sigma I_D$

$$\boxed{\text{if } \sigma \to 0 \quad \xrightarrow[\text{get K-means}]{} \quad z_{kn}}$$

2. Compute the marginal likelihood (cost).

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

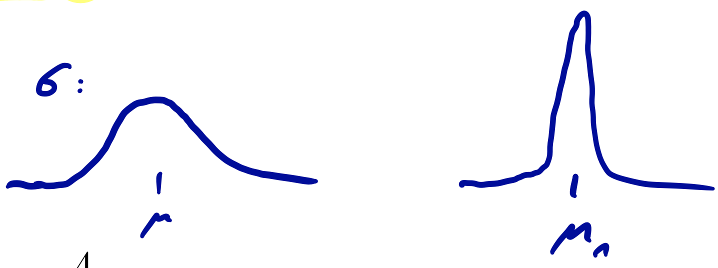
3. M-step: Update $\left(\boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}, \pi_k^{(t+1)}.\right) = \theta^{(t+1)}$

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

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

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

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



computing q_{kn}

E-step

M-step: new μ, Σ, π

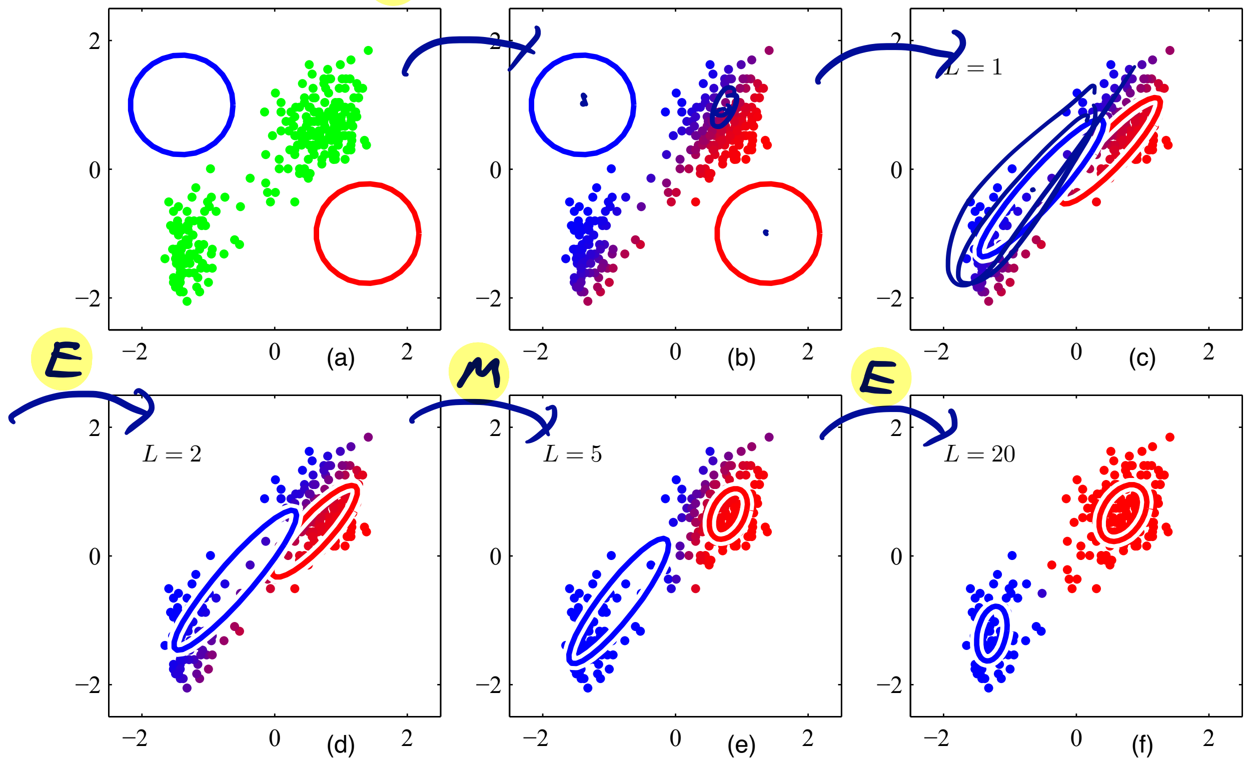
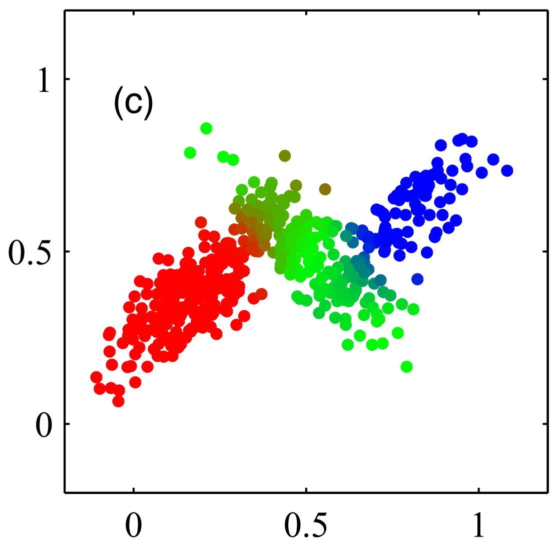


Figure 1: EM algorithm for GMM

Posterior distribution

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

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



EM in general

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

The EM algorithm can be compactly written as follows:

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

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.