annotated version

Machine Learning Course - CS-433

Expectation-Maximization Algorithm

Dec 2, 2021

changes by Martin Jaggi 2021, 2020, 2019, changes by Rüdiger Urbanke 2018, changes by Martin Jaggi 2017, 2016

Last updated on: November 30, 2021

EPFL

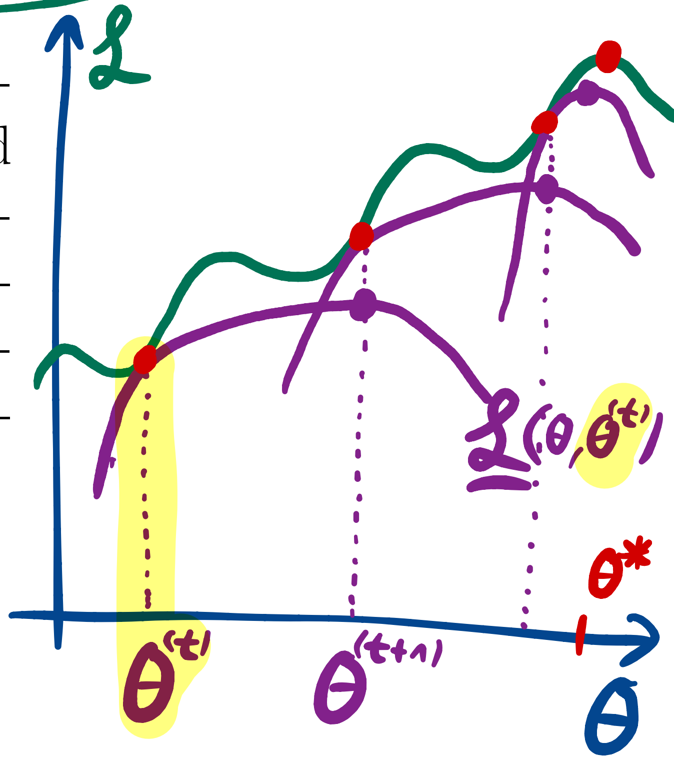
Motivation

$\theta = (\mu, \Sigma, \pi)$

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

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

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

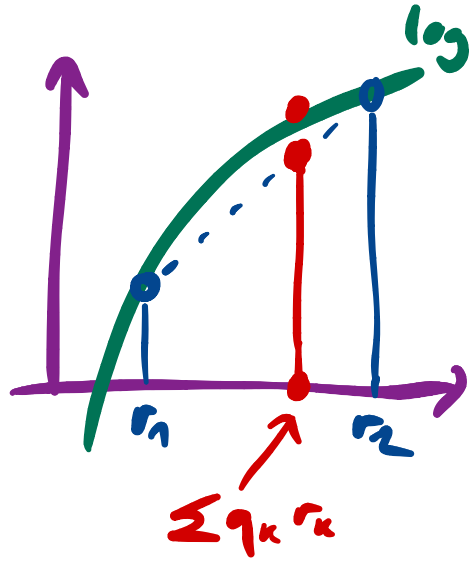
$\mathcal{L}(\boldsymbol{\theta}) \geq \underline{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t)})$ and
$\mathcal{L}(\boldsymbol{\theta}^{(t)}) = \underline{\mathcal{L}}(\boldsymbol{\theta}^{(t)}, \boldsymbol{\theta}^{(t)})$.

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

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

⟺ convexity of $-\log$

Concavity of log



Given non-negative weights q s.t. $\sum_k q_k = 1$, the following holds for any $r_k > 0$: ⟺ Jensen's Inequality

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

The expectation step

lower bound on $\mathcal{L}$

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

(b) 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.

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

$$= \sum_{k=1}^{K} q_{kn}^{(t)} \log \bigcirc$$

$$= \sum_{k=1}^{K} \frac{\pi_k^{(t)} \mathcal{N}(x_n | \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'} \pi_{k'} \mathcal{N}(\dots)} \log \frac{\cancel{\pi_k \mathcal{N}(x, \dots)}}{\cancel{\pi_k \mathcal{N}(x \dots)} \,/\, \sum_{k'} \pi_{k'} \mathcal{N}(x \dots)}$$

$$= \log \sum_{k'} \pi_{k'} \mathcal{N}(x_n, \mu_k, \Sigma_k) \cdot 1$$

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

$\mathcal{N}(x|\mu,\Sigma) = e^{-(x-\mu)^T \Sigma^{-1} (x-\mu)}$

The maximization step

constant in θ

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

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

$$= \log \frac{\pi_k \mathcal{N}(x_n, \dots)}{q_{kn}^{(t)}}$$

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

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

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

$v = x - \mu$

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

$$\nabla_{\Sigma_k^{-1}} \mathcal{L}(\theta, \theta^{(t)}) \overset{!}{=} 0$$

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:

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

!!

$$\mathcal{L}_n(\theta, \theta^{(t)}) + \beta\left(\sum \pi_k - 1\right)$$

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

Summary of EM for GMM

and k-means as special case

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 $q_{kn}^{(t)}$:

at $\theta^{(t)}$

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

$\approx \exp \boxed{-\frac{\|x_n - \mu_k\|^2}{\sigma^2}}$

k-means $\parallel$ z_{kn}

$\xrightarrow{\sigma^2 \to 0}$ $\begin{cases} 1 & k \text{ closest to } x_n \\ 0 & \text{otherwise} \end{cases}$

$\boxed{= \text{k-means assignment}}$

2. Compute the marginal likelihood (cost).

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

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

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

$\leftarrow$ mean of cluster

$\theta^{t+1} =$

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

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

$\leftarrow$ # points assigned to cluster k

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 $\sigma^2 \to 0$.

σ width

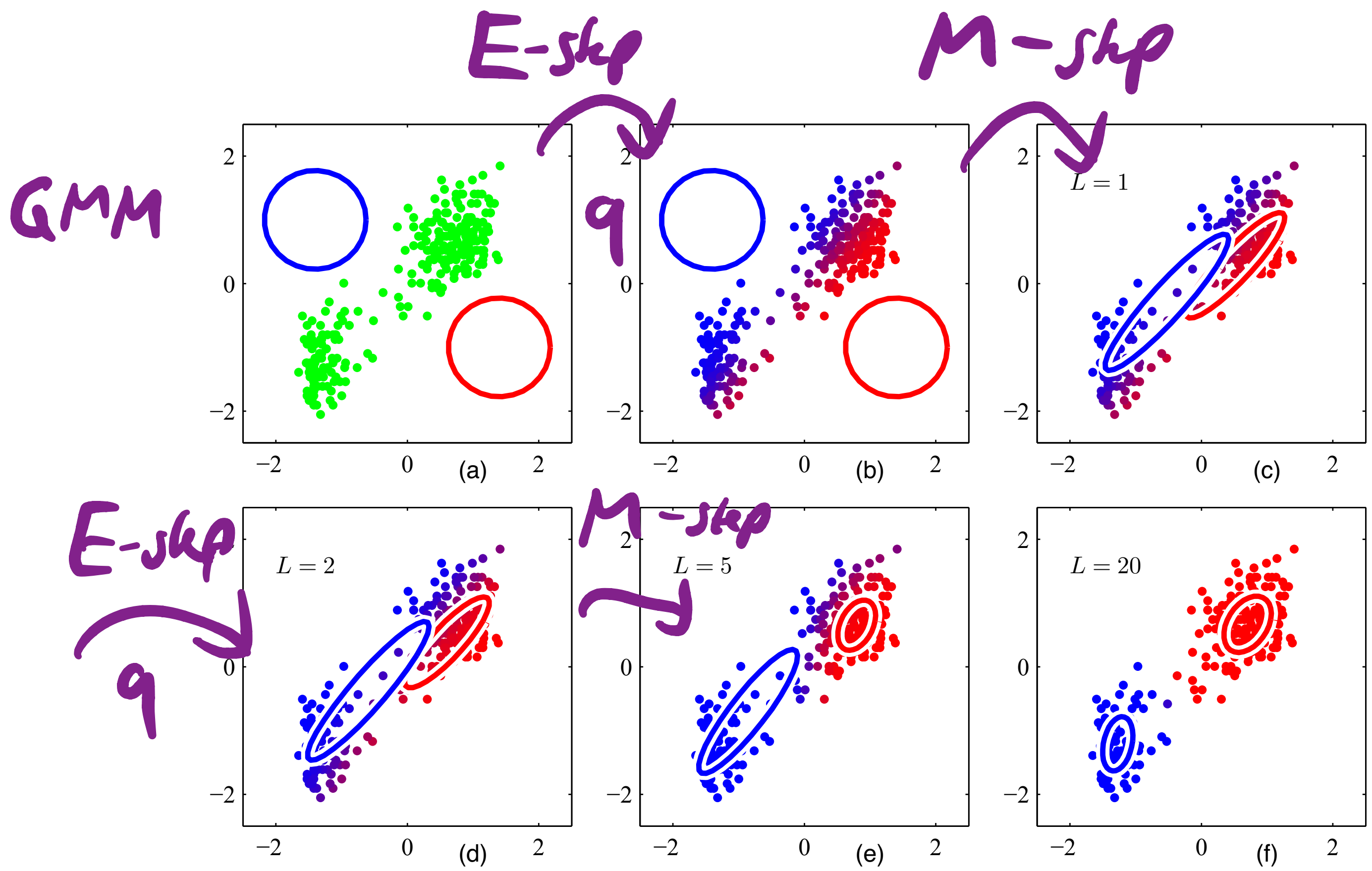
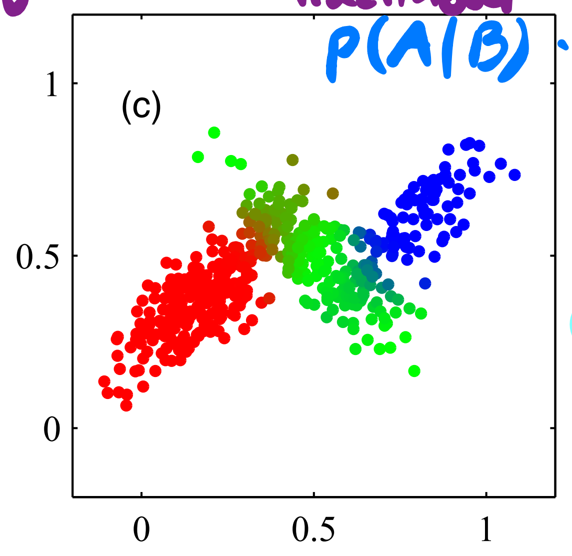


Figure 1: EM algorithm for GMM

Posterior distribution

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

$$\underbrace{p(\mathbf{x}_n, z_n|\boldsymbol{\theta})}_{\text{joint}} = \underbrace{p(\mathbf{x}_n|z_n, \boldsymbol{\theta})}_{\text{likelihood}} \underbrace{p(z_n|\boldsymbol{\theta})}_{\text{prior}} = \underbrace{p(z_n|\mathbf{x}_n, \boldsymbol{\theta})}_{\text{posterior}} \underbrace{p(\mathbf{x}_n|\boldsymbol{\theta})}_{\text{marginal likelihood}}$$



$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Bayes Rule

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

$$= \frac{\overbrace{p(z_n=k|\theta)}^{\pi_k} \cdot \overbrace{p(x_n|z_n,\theta)}^{\mathcal{N}(x_n,\mu_k,\Sigma_k)}}{\sum \underbrace{p(z_n=k)}_{\pi_k} \cdot \underbrace{p(x|z,\theta)}_{\mathcal{N}(x_n|\mu_k,\Sigma_k)}}$$

$$= q_{nk}$$

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:

Expectation over z

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

Maximization

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.