annotated version

Machine Learning Course - CS-433

Expectation-Maximization Algorithm

Nov 26, 2020

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

Last updated on: November 24, 2020

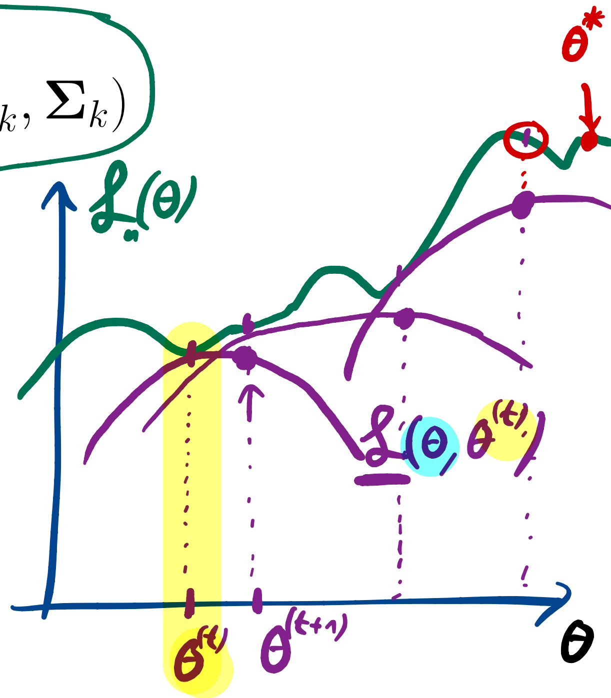
Motivation

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

$\mathcal{L}_n(\theta)$

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

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



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

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

- lower bound
- equality if $\theta = \theta^{(t)}$

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

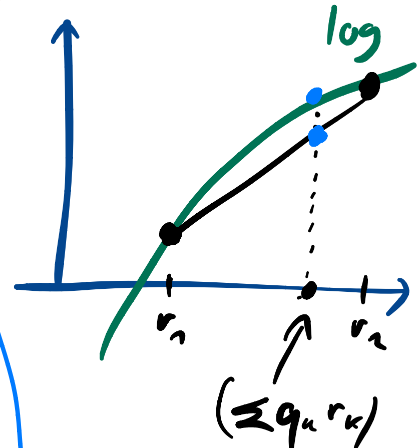
$\Leftarrow$ Convexity of $-\log$

Concavity of log

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

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

$\Leftarrow$ Jensen's Inequality



The expectation step

lower bound to $\mathcal{L}_n(\theta)$

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

with equality when,

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

This is not a coincidence.

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

$$= \underbrace{\sum_{k=1}^{K} \underbrace{\boxed{\frac{\pi_k \mathcal{N}(x_n, \ldots)}{\sum_{k'} \pi_{k'} \mathcal{N}(x_n, \ldots)}}}_{q_{kn}}}_{=1} \log \frac{\cancel{\pi_k \mathcal{N}(x_n, \ldots)}}{\frac{\cancel{\pi_k \mathcal{N}(x_n \ldots)}}{\sum_{k'} \pi_{k'} \mathcal{N}(x_n, \ldots)}}$$

$$= \log \sum_{k'=1}^{K} \pi_{k'} \mathcal{N}(x_n | \mu_{k'}, \Sigma_{k'})$$

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

- lower bound ✓
- coincides at $\theta = \theta^{(t)}$ ✓

The maximization step

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

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

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

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

independent of θ

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

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

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

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

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

$v \, v^\top$ = rank one matrix

$v = x_n - \mu_k$

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:

constraint

want: $\sum_k \pi_k = 1$

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

maximize $\mathcal{L}_n + \beta\left(\sum_k \pi_k - 1\right)$

(unconstrained)

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

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

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

$$\approx \frac{\exp\left(\frac{-\|x_n - \mu_k\|^2}{\sigma^2}\right)}{\sum_{k=1}^{K} \exp(-\|\ldots\|/\sigma^2)}$$

$\xrightarrow{\sigma^2 \to 0}$ $\begin{cases} 1 & \text{closest } k \\ 0 & \text{others} \end{cases}$ $\approx$ k-means assignment

$q_{kn} \approx z_{kn}$

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 $\boldsymbol{\mu}_k^{(t+1)}, \boldsymbol{\Sigma}_k^{(t+1)}, \pi_k^{(t+1)}$.

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

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

$$\pi_k^{(t+1)} := \frac{1}{N} \sum_n q_{kn}^{(t)} \quad \leftarrow \text{\# points assigned to } k$$

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

width

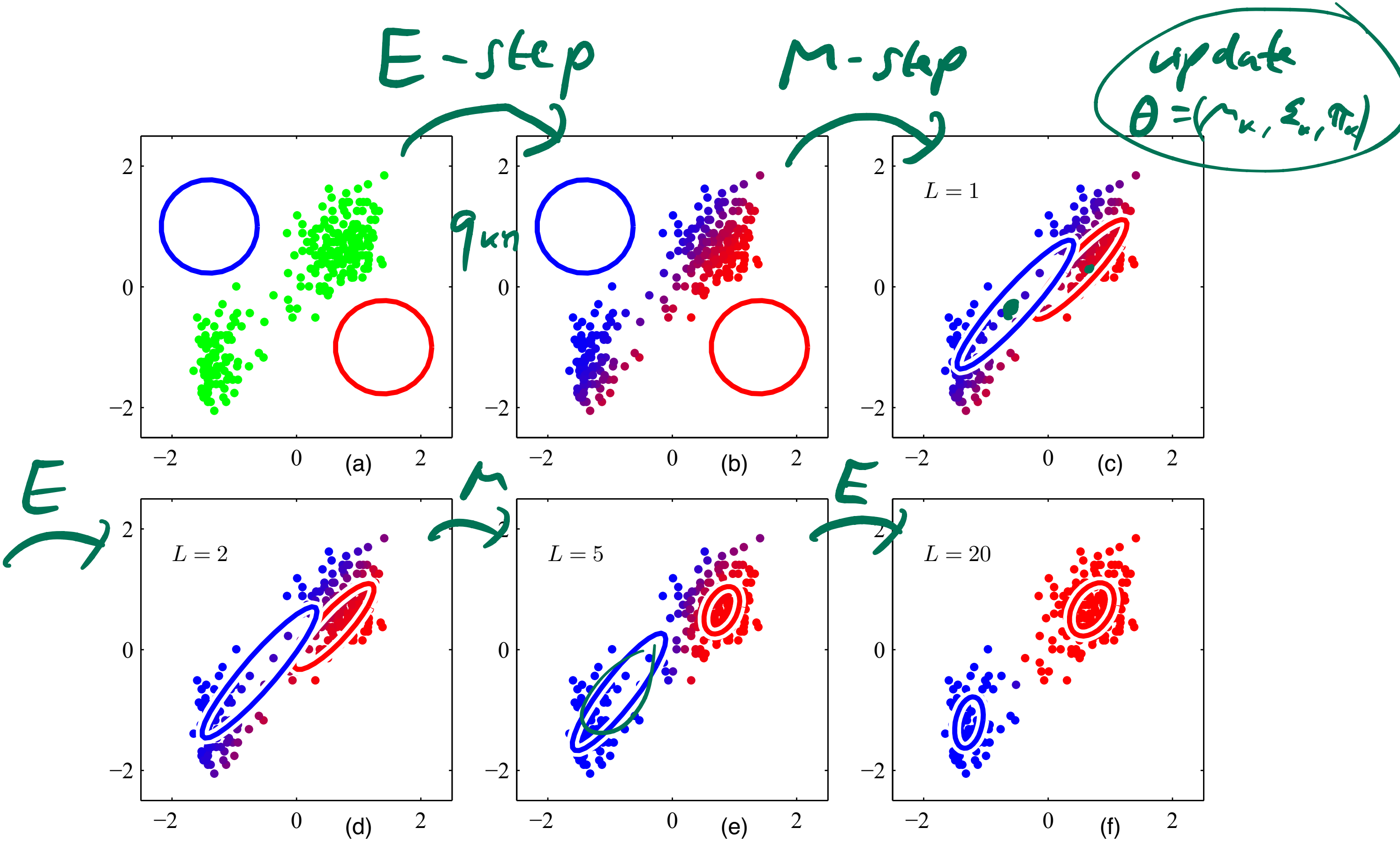


Figure 1: EM algorithm for GMM

Posterior distribution

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

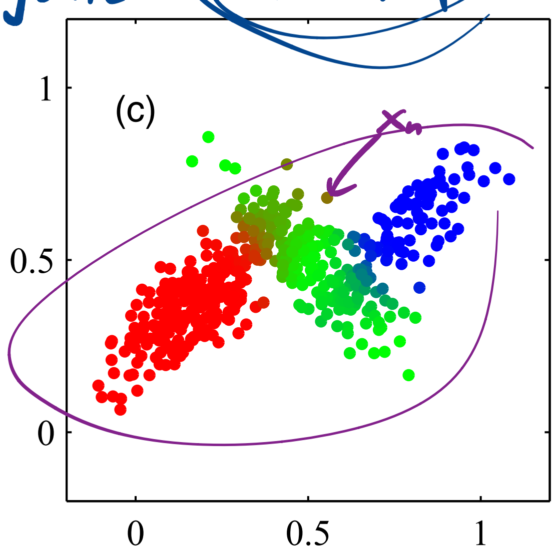
posterior

Bayes rule

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

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

joint · likelihood · prior · posterior · marginal likelihood



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

$$= \frac{\overbrace{p(z_n = k|\theta)}^{\pi_k}\, p(x_n|z_n, \theta)}{\sum_{k=1}^{K} \underbrace{p(z_n = k)}_{\pi_k}\, p(x|z_n, \theta)}$$

$$= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(x_n| \ldots)} =: q_{kn}$$

EM in general

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

The EM algorithm can be compactly written as follows:

hidden variables z

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

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.