

Gaussian Mixture Model

He He

Slides based on Lecture 13b from David Rosenberg's course materials
(<https://github.com/davidrosenberg/mlcourse>)

CDS, NYU

April 27, 2021

Latent Variable Models

General Latent Variable Model

- Two sets of random variables: z and x .
- z consists of unobserved **hidden variables**.
- x consists of **observed variables**.
- Joint probability model parameterized by $\theta \in \Theta$:

$$p(x, z \mid \theta)$$

Definition

A **latent variable model** is a probability model for which certain variables are never observed.

e.g. The Gaussian mixture model is a latent variable model.

Complete and Incomplete Data

- Suppose we observe some data (x_1, \dots, x_n) .
- To simplify notation, take x to represent the entire dataset

$$x = (x_1, \dots, x_n),$$

and z to represent the corresponding unobserved variables

$$z = (z_1, \dots, z_n).$$

- An observation of x is called an **incomplete data set**.
- An observation (x, z) is called a **complete data set**.

Our Objectives

- **Learning problem:** Given incomplete dataset x , find MLE

$$\hat{\theta} = \arg \max_{\theta} p(x | \theta).$$

- **Inference problem:** Given x , find conditional distribution over z :

$$p(z | x, \theta).$$

- For Gaussian mixture model, learning is hard, inference is easy.
- For more complicated models, inference can also be hard. (See DSGA-1005)

Log-Likelihood and Terminology

- Note that

$$\arg \max_{\theta} p(x \mid \theta) = \arg \max_{\theta} [\log p(x \mid \theta)].$$

- Often easier to work with this “**log-likelihood**”.
- We often call $p(x)$ the **marginal likelihood**,
 - because it is $p(x, z)$ with z “marginalized out”:

$$p(x) = \sum_z p(x, z)$$

- We often call $p(x, z)$ the **joint**. (for “joint distribution”)
- Similarly, $\log p(x)$ is the **marginal log-likelihood**.

EM Algorithm

Intuition

Problem: marginal log-likelihood $\log p(x; \theta)$ is hard to optimize (observing only x)

Observation: complete data log-likelihood $\log p(x, z; \theta)$ is easy to optimize (observing both x and z)

Idea: guess a distribution of the latent variables $q(z)$ (soft assignments)

Maximize the **expected complete data log-likelihood**:

$$\max_{\theta} \sum_{z \in \mathcal{Z}} q(z) \log p(x, z; \theta)$$

EM assumption: the expected complete data log-likelihood is easy to optimize

Why should this work?

Math Prerequisites

Jensen's Inequality

Theorem (Jensen's Inequality)

If $f : \mathbb{R} \rightarrow \mathbb{R}$ is a **convex** function, and x is a random variable, then

$$\mathbb{E}f(x) \geq f(\mathbb{E}x).$$

Moreover, if f is **strictly convex**, then equality implies that $x = \mathbb{E}x$ with probability 1 (i.e. x is a constant).

- e.g. $f(x) = x^2$ is convex. So $\mathbb{E}x^2 \geq (\mathbb{E}x)^2$. Thus

$$\text{Var}(x) = \mathbb{E}x^2 - (\mathbb{E}x)^2 \geq 0.$$

Kullback-Leibler Divergence

- Let $p(x)$ and $q(x)$ be probability mass functions (PMFs) on \mathcal{X} .
- How can we measure how “different” p and q are?
- The **Kullback-Leibler** or “**KL**” **Divergence** is defined by

$$\text{KL}(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

(Assumes $q(x) = 0$ implies $p(x) = 0$.)

- Can also write this as

$$\text{KL}(p\|q) = \mathbb{E}_{x \sim p} \log \frac{p(x)}{q(x)}.$$

Gibbs Inequality ($\text{KL}(p\|q) \geq 0$ and $\text{KL}(p\|p) = 0$)

Theorem (Gibbs Inequality)

Let $p(x)$ and $q(x)$ be PMFs on \mathcal{X} . Then

$$\text{KL}(p\|q) \geq 0,$$

with equality iff $p(x) = q(x)$ for all $x \in \mathcal{X}$.

- KL divergence measures the “distance” between distributions.
- Note:
 - KL divergence **not a metric**.
 - KL divergence is **not symmetric**.

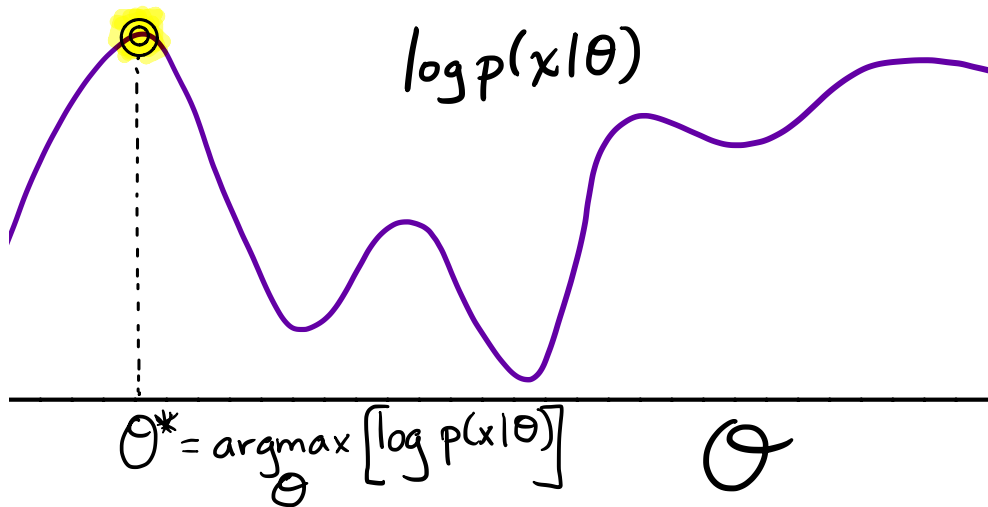
Gibbs Inequality: Proof

$$\begin{aligned}\mathrm{KL}(p\|q) &= \mathbb{E}_p \left[-\log \left(\frac{q(x)}{p(x)} \right) \right] \\ &\geq -\log \left[\mathbb{E}_p \left(\frac{q(x)}{p(x)} \right) \right] \quad (\text{Jensen's}) \\ &= -\log \left[\sum_{\{x|p(x)>0\}} p(x) \frac{q(x)}{p(x)} \right] \\ &= -\log \left[\sum_{x \in \mathcal{X}} q(x) \right] \\ &= -\log 1 = 0.\end{aligned}$$

- Since $-\log$ is strictly convex, we have strict equality iff $q(x)/p(x)$ is a constant, which implies $q = p$.

The ELBO: Family of Lower Bounds on $\log p(x | \theta)$

The Maximum Likelihood Estimator



Lower bound of the marginal log-likelihood

$$\begin{aligned}\log p(x; \theta) &= \log \sum_{z \in \mathcal{Z}} p(x, z; \theta) \\ &= \log \sum_{z \in \mathcal{Z}} q(z) \frac{p(x, z; \theta)}{q(z)} \\ &\geq \sum_{z \in \mathcal{Z}} q(z) \log \frac{p(x, z; \theta)}{q(z)} \\ &\stackrel{\text{def}}{=} \mathcal{L}(q, \theta)\end{aligned}$$

- **Evidence:** $\log p(x; \theta)$
- **Evidence lower bound (ELBO):** $\mathcal{L}(q, \theta)$
- q : chosen to be a family of tractable distributions
- Idea: *maximize the ELBO* instead of $\log p(x; \theta)$

MLE, EM, and the ELBO

- The MLE is defined as a maximum over θ :

$$\hat{\theta}_{\text{MLE}} = \arg \max_{\theta} [\log p(x | \theta)].$$

- For any PMF $q(z)$, we have a lower bound on the marginal log-likelihood

$$\log p(x | \theta) \geq \mathcal{L}(q, \theta).$$

- In EM algorithm, we maximize the lower bound (ELBO) over θ and q :

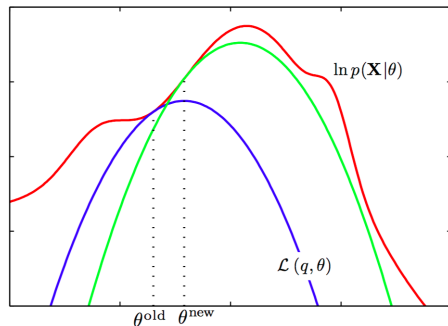
$$\hat{\theta}_{\text{EM}} \approx \arg \max_{\theta} \left[\max_q \mathcal{L}(q, \theta) \right]$$

- In EM algorithm, q ranges over all distributions on z .

EM: Coordinate Ascent on Lower Bound

- Choose sequence of q 's and θ 's by “**coordinate ascent**” on $\mathcal{L}(q, \theta)$.
- EM Algorithm (high level):
 - ① Choose initial θ^{old} .
 - ② Let $q^* = \arg \max_q \mathcal{L}(q, \theta^{\text{old}})$
 - ③ Let $\theta^{\text{new}} = \arg \max_{\theta} \mathcal{L}(q^*, \theta)$.
 - ④ Go to step 2, until converged.
- Will show: $p(x | \theta^{\text{new}}) \geq p(x | \theta^{\text{old}})$
- Get sequence of θ 's with monotonically increasing likelihood.

EM: Coordinate Ascent on Lower Bound



- 1 Start at θ^{old} .
- 2 Find q giving best lower bound at $\theta^{\text{old}} \implies \mathcal{L}(q, \theta)$.
- 3 $\theta^{\text{new}} = \arg \max_{\theta} \mathcal{L}(q, \theta)$.

From Bishop's *Pattern recognition and machine learning*, Figure 9.14.

Justification for maximizing ELBO

$$\begin{aligned}\mathcal{L}(q, \theta) &= \sum_{z \in \mathcal{Z}} q(z) \log \frac{p(x, z; \theta)}{q(z)} \\&= \sum_{z \in \mathcal{Z}} q(z) \log \frac{p(z | x; \theta) p(x; \theta)}{q(z)} \\&= - \sum_{z \in \mathcal{Z}} q(z) \log \frac{q(z)}{p(z | x; \theta)} + \sum_{z \in \mathcal{Z}} q(z) \log p(x; \theta) \\&= -\text{KL}(q(z) \| p(z | x; \theta)) + \underbrace{\log p(x; \theta)}_{\text{evidence}}\end{aligned}$$

- **KL divergence:** measures “distance” between two distributions (not symmetric!)
- $\text{KL}(q \| p) \geq 0$ with equality iff $q(z) = p(z | x)$.
- $\text{ELBO} = \text{evidence} - \text{KL} \leq \text{evidence}$

Maximizing over q for fixed θ .

- Find q maximizing

$$\mathcal{L}(q, \theta) = -\text{KL}[q(z), p(z | x, \theta)] + \underbrace{\log p(x | \theta)}_{\text{no } q \text{ here}}$$

- Recall $\text{KL}(p \| q) \geq 0$, and $\text{KL}(p \| p) = 0$.

- Best q is $q^*(z) = p(z | x, \theta)$ and

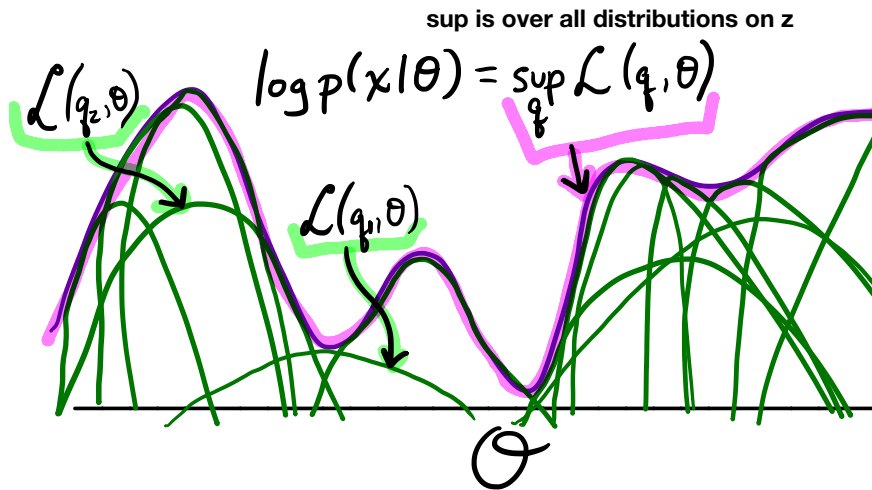
$$\mathcal{L}(q^*, \theta) = -\underbrace{\text{KL}[p(z | x, \theta), p(z | x, \theta)]}_{=0} + \log p(x | \theta)$$

- Summary:

$$\log p(x | \theta) = \sup_q \mathcal{L}(q, \theta) \quad \forall \theta$$

- For any θ , **sup is attained** at $q(z) = p(z | x, \theta)$.

Marginal Log-Likelihood **IS** the Supremum over Lower Bounds



Summary

Latent variable models: clustering, latent structure, missing labels etc.

Parameter estimation: maximum marginal log-likelihood

Challenge: directly maximize the **evidence** $\log p(x; \theta)$ is hard

Solution: maximize the **evidence lower bound**:

$$\text{ELBO} = \mathcal{L}(q, \theta) = -\text{KL}(q(z) \| p(z | x; \theta)) + \log p(x; \theta)$$

Why does it work?

$$\begin{aligned} q^*(z) &= p(z | x; \theta) \quad \forall \theta \in \Theta \\ \mathcal{L}(q^*, \theta^*) &= \max_{\theta} \log p(x; \theta) \end{aligned}$$

EM algorithm

Coordinate ascent on $\mathcal{L}(q, \theta)$

- 1 Random initialization: $\theta^{\text{old}} \leftarrow \theta_0$
- 2 Repeat until convergence
 - i $q(z) \leftarrow \arg \max_q \mathcal{L}(q, \theta^{\text{old}})$

Expectation (the E-step): $q^*(z) = p(z \mid x; \theta^{\text{old}})$
 $J(\theta) = \mathcal{L}(q^*, \theta)$

ii $\theta^{\text{new}} \leftarrow \arg \max_{\theta} \mathcal{L}(q^*, \theta)$

Maximization (the M-step): $\theta^{\text{new}} \leftarrow \arg \max_{\theta} J(\theta)$

1 Expectation Step

- Let $q^*(z) = p(z \mid x, \theta^{\text{old}})$. [q^* gives best lower bound at θ^{old}]
- Let

$$J(\theta) := \mathcal{L}(q^*, \theta) = \underbrace{\sum_z q^*(z) \log \left(\frac{p(x, z \mid \theta)}{q^*(z)} \right)}_{\text{expectation w.r.t. } z \sim q^*(z)}$$

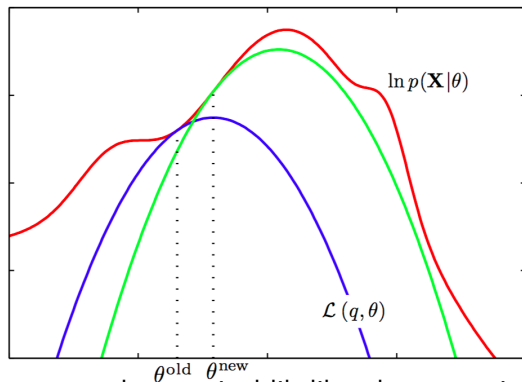
2 Maximization Step

$$\theta^{\text{new}} = \arg \max_{\theta} J(\theta).$$

[Equivalent to maximizing expected complete log-likelihood.]

EM puts no constraint on q in the E-step and assumes the M-step is easy. In general, both steps can be hard.

[discussion] Monotonically increasing likelihood



Exercise: prove that EM increases the marginal likelihood monotonically

$$\log p(x; \theta^{\text{new}}) \geq \log p(x; \theta^{\text{old}}) .$$

Does EM converge to a global maximum?

Variations on EM

EM Gives Us Two New Problems

- The “E” Step: Computing

$$J(\theta) := \mathcal{L}(q^*, \theta) = \sum_z q^*(z) \log \left(\frac{p(x, z | \theta)}{q^*(z)} \right)$$

- The “M” Step: Computing

$$\theta^{\text{new}} = \arg \max_{\theta} J(\theta).$$

- Either of these can be too hard to do in practice.

Generalized EM (GEM)

- Addresses the problem of a difficult “M” step.
- Rather than finding

$$\theta^{\text{new}} = \arg \max_{\theta} J(\theta),$$

find **any** θ^{new} for which

$$J(\theta^{\text{new}}) > J(\theta^{\text{old}}).$$

- Can use a standard nonlinear optimization strategy
 - e.g. take a gradient step on J .
- We still get monotonically increasing likelihood.

EM and More General Variational Methods

- Suppose “E” step is difficult:
 - Hard to take expectation w.r.t. $q^*(z) = p(z | x, \theta^{\text{old}})$.
- Solution: Restrict to distributions \mathcal{Q} that are easy to work with.
- Lower bound now looser:

$$q^* = \arg \min_{q \in \mathcal{Q}} \text{KL}[q(z), p(z | x, \theta^{\text{old}})]$$