13. Expectation Maximization (EM)

Dongwoo Kim

dongwookim.ac.kr

CSED515 - 2023 Spring

Estimation with Latent Variables

When there are missing data or latent variables, denoted by z, MLE seeks to find θ maximizing the marginal likelihood of the observed data x:

$$p(x \mid \theta) = \int p(x, z \mid \theta) dz \ .$$

As such, MLE or MAP often require the computationally intractable marginalization or maximization. Variational inference is a family of techniques to approximate the marginalization or maximization, e.g.,

- Belief propagation
- Expectation-maximization
- Mean field approximation
- ...

Outline

Analysis and generalization of EM algorithm:

- Mathematical preliminaries
  - Jensen's inequality and Gibb's inequality
  - Entropy and mutual information
- Expectation-Maximization algorithms
  - The monotonicity property of EM algorithm
  - Generalizations of EM algorithm
  - EM algorithm for exponential family

Convex Set and Function

- A set $C \subset \mathbb{R}^d$ is convex if

$$\lambda x + (1-\lambda) y \in C \ , \quad \forall x, y \in C \text{ and } \forall \lambda \in [0,1].$$

- For a convex set $C \subset \mathbb{R}^d$, a function $f : C \mapsto \mathbb{R}$ is convex if

$$f(\lambda x + (1-\lambda) y) \leq \lambda f(x) + (1-\lambda) f(y) \ , \quad \forall x, y \in C \text{ and } \forall \lambda \in [0,1].$$

Jensen's Inequality

Theorem (Jensen's inequality for random variables)

For a convex set C, if function $f : C \mapsto \mathbb{R}$ is convex and X is a random vector on C, then

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]) \,.$$

In case of concave f, we have $\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$.

Proof of Jensen's Inequality

For simplicity, consider discrete random vector X with $p_i = p(X = x_i)$ for $\{x_i\}_{i\in[N]} \subset C$. We prove $\sum_{i\in[N]} p_i f(x_i) \geq f(\sum_{i\in[N]} p_i x_i)$ by recursion:

$$
\begin{aligned}
& f\left(\sum_{i\in[N]} p_i x_i\right) = f\left(p_1 x_1 + (1-p_1)\left(\frac{\sum_{i=2}^N p_i x_i}{1-p_1}\right)\right) \\
& \leq p_1 f(x_1) + (1-p_1) f\left(\frac{\sum_{i=2}^N p_i x_i}{1-p_1}\right) \\
& = p_1 f(x_1) + (1-p_1) f\left(\frac{p_2}{1-p_1} x_2 + \left(\frac{1-\sum_{i=1}^2 p_i}{1-p_1}\right)\left(\frac{\sum_{i=3}^N p_i x_i}{1-\sum_{i=1}^2 p_i}\right)\right) \\
& \leq p_1 f(x_1) + (1-p_1)\left(\left(\frac{p_2}{1-p_1}\right) f(x_2) + \left(\frac{1-\sum_{i=1}^2 p_i}{1-p_1}\right) f\left(\frac{\sum_{i=3}^N p_i x_i}{1-\sum_{i=1}^2 p_i}\right)\right) \\
& = p_1 f(x_1) + p_2 f(x_2) \left(1 - \sum_{i=1}^2 p_i\right) f\left(\frac{\sum_{i=3}^N p_i x_i}{1-\sum_{i=1}^2 p_i}\right) \ldots
\end{aligned}
$$

Information and Entropy

- Information $I(X)$ of random variable X is defined as

$$I(X) := -\log p(X) \ ,$$

  which is itself a random variable, and quantifies the surprise or uncertainty of the realization of X.

- Entropy $H(X)$ of random variable X is defined as the expected value of information:

$$H(X) := \mathbb{E}[I(X)] = -\sum_{x \in \mathcal{X}} p(x) \log_b p(x) \ ,$$

  which measures the uncertainty of X w.r.t. base $b > 0$, and $\mathcal{X}$ is the set of all possible values of X.

In fact, those concepts were developed in the information theory to study communication system. The entropy $H(X)$ can be interpreted as the minimum bits to express data X.

Entropy and Relative Entropy

- Entropy is a measure of uncertainty of a random variable, defined by:

$$H(X) := \mathbb{E}[I(X)] = -\sum_{x \in \mathcal{X}} p(x) \log p(x) .$$

- Kullback-Leibler divergence is a measure of relative entropy of distribution p to reference distribution q such that p is absolutely continuous w.r.t. q, i.e., $q(x) = 0$ implies $p(x) = 0$, defined by:

$$\mathrm{KL}(p \| q) := \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} ,$$

where we use the convention of $0 \log(0/0) = 0$.

Gibb's Inequality

Theorem (Gibb's Inequality)

For any distributions p, q such that $p \ll q$, i.e., p is absolutely continuous w.r.t. q,

$$KL(p\|q) \geq 0 \ ,$$

where the equality holds iff $p = q$.

Proof) Consider discrete distributions $\{p_i\}, \{q_i\}$.

$$\begin{aligned}
\mathrm{KL}(p\|q) = \sum_i p_i \log \frac{p_i}{q_i} &= -\sum_i p_i \log \frac{q_i}{p_i} \\
&\geq -\log\left(\sum_i p_i \frac{q_i}{p_i}\right) \qquad \text{(Jensen's ineq.)} \\
&= -\log\left(\sum_i q_i\right) = 0 \ .
\end{aligned}$$

Gibb's Inequality: Proof of the Equality

In order to find the "distribution" p which minimizes $\mathrm{KL}(p\|q)$, we consider Lagrangian

$$\mathcal{F}(p,\lambda) = \mathrm{KL}(p\|q) + \lambda\left(1 - \sum_i p_i\right) = \sum_i p_i \log\frac{p_i}{q_i} + \lambda\left(1 - \sum_i p_i\right) .$$

Then, the minimal p must have λ verifying:

$$\frac{\partial \mathcal{F}}{\partial p_i} = \log p_i - \log q_i + 1 - \lambda = 0 ,$$

which implies $p_i = q_i \exp(\lambda - 1)$ for each i. Using $\sum_i p_i = 1 = \sum_i q_i \exp(\lambda - 1)$, it follows that $\lambda = 1$. Hence, the minimal p should be identical to q, and $\mathrm{KL}(p\|q) = 0$ on such choice of p.

Outline

Analysis and generalization of EM algorithm:

- Mathematical preliminaries
  - Jensen's inequality and Gibb's inequality
  - Entropy and mutual information
- **Expectation-Maximization algorithms**
  - The monotonicity property of EM algorithm
  - Generalizations of EM algorithm
  - EM algorithm for exponential family

A Lower Bound on the Log-Likelihood (1)

The log-likelihood of model parameter θ given observation x is:

$$\begin{aligned}\mathcal{L}(\theta) &= \log p(x \mid \theta) \\ &= \log \int p(x, z \mid \theta) dz\ ,\end{aligned}$$

where we marginalize out the latent variables z in the second equality.
For any distribution $q(z)$ of the latent variables z, we have

$$\begin{aligned}\mathcal{L}(\theta) &= \log \left(\int q(z) \frac{p(x, z \mid \theta)}{q(z)} dz \right) \\ &\geq \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz \qquad \text{(Jensen's ineq.)}\ .\end{aligned}$$

A Lower Bound on the Log-Likelihood (2)

Denote the lower bound by $\mathcal{F}(q, \theta)$:

$$\begin{aligned}\mathcal{F}(q, \theta) &:= \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz \\ &= \int q(z) \log p(x, z \mid \theta) dz + H(q) \quad \text{(Def. of entropy)} .\end{aligned}$$

where $H(q)$ is the entropy of q.
One can design an EM algorithm using this lower bound:

- E-step: Maximize $\mathcal{F}(q, \theta)$ over q for tighter lower bound
- M-step: Maximize $\mathcal{F}(q, \theta)$ over θ to update estimates of θ.

EM Algorithm with max-max Interpretation

(for $k = 1, 2, \ldots$)

- E-step: Optimize $\mathcal{F}(q, \theta)$ w.r.t. the distribution q of latent variable z given parameters $\theta^{(k)}$, i.e.,

$$q^{(k+1)} = \arg\max_{q} \ \ \mathcal{F}(q, \theta^{(k)}) \ .$$

- M-step: Maximize $\mathcal{F}(q, \theta)$ w.r.t. the parameters θ given the distribution $q^{(k+1)}$ of latent variable z, i.e.,

$$\begin{aligned} \theta^{(k+1)} &= \arg\max_{\theta} \ \ \mathcal{F}(q^{(k+1)}, \theta) \\ &= \arg\max_{\theta} \ \ \int q^{(k+1)}(z) \log p(x, z \mid \theta) dz \ , \end{aligned}$$

where $p(x, z \mid \theta)$ is the complete-data log-likelihood.

Monotonicity of EM Algorithm

The difference between the log-likelihood and the lower bound is:

$$\begin{aligned}
\mathcal{L}(\theta) - \mathcal{F}(q,\theta) &= \log p(x \mid \theta) - \int q(z) \log \left(\frac{p(x, z \mid \theta)}{q(z)} \right) dz \\
&= \log p(x \mid \theta) - \int q(z) \log \left(\frac{p(z \mid x, \theta) p(x \mid \theta)}{q(z)} \right) dz \\
&= - \int q(z) \log \left(\frac{p(z \mid x, \theta)}{q(z)} \right) dz \\
&= \mathrm{KL}(q(\cdot) \| p(\cdot \mid x, \theta)) \, ,
\end{aligned}$$

which is zero only if $q(z) = p(z \mid x, \theta)$ (Gibb's ineq.). This is what E-step finds. Hence,

$$\mathcal{L}(\theta^{(k)}) \underset{\text{E-step}}{=} \mathcal{F}(q^{(k+1)}, \theta^{(k)}) \underset{\text{M-step}}{\leq} \mathcal{F}(q^{(k+1)}, \theta^{(k+1)}) \underset{\text{Jensen}}{\leq} \mathcal{L}(\theta^{(k+1)}) \, .$$

EM Algorithm

The EM algorithm seeks to find the MLE by iteratively applying:
(for $k = 1, 2, \ldots$)

- E-step: Define $\mathcal{Q}(\theta; \theta^{(k)})$ as the expectation of complete-data log-likelihood w.r.t. z given x and $\theta^{(k)}$:

$$\begin{aligned}
\mathcal{Q}(\theta; \theta^{(k)}) &:= \mathbb{E}_{z|x,\theta^{(k)}}[\log p(x, z \mid \theta)] \\
&= \int p(z \mid x, \theta^{(k)}) \log p(x, z \mid \theta) dz \ .
\end{aligned}$$

- M-step: Find the parameters that maximize:

$$\begin{aligned}
\theta^{(k+1)} &:= \arg\max_{\theta} \mathcal{Q}(\theta; \theta^{(k)}) \\
&= \arg\max_{\theta} \mathcal{F}(q, \theta) - H(q) \\
&\quad \text{(with the choice of } q(z) = p(z|x, \theta^{(k)})) \ ,
\end{aligned}$$

where the term $H(q)$ is ignored since $H(q)$ is constant w.r.t. θ.

Additional Slides: Exponential Family

The exponential family is a family probability distribution functions each of which has a special form given by

$$p(x \mid \theta) = h(x)g(\eta)\exp\left(\eta^\top u(x)\right) ,$$

where $\eta = \eta(\theta)$ is a function of θ, and $h(x), u(x)$ and $g(\eta)$ are known. The function $g(\eta)$ normalizes the distribution so that

$$g(\eta)\int h(x)\exp\left(\eta^\top u(x)\right) dx = 1 .$$

where the integration is replaced with sum for the case of discrete x.

Example of Exponential Family: Bernoulli

Consider a Bernoulli variable x with mean $\theta \in (0, 1)$ of which distribution can be expressed as follows:

$$\begin{aligned} p(x \mid \theta) = \mathrm{Bern}(x \mid \theta) &= \theta^x (1-\theta)^{1-x} \\ &= \exp\left(x \log\theta + (1-x)\log(1-\theta)\right) \\ &= (1-\theta)\exp\left(\log\left(\frac{\theta}{1-\theta}\right) x\right) \ , \end{aligned}$$

which implies that the Bernoulli variable is exponential family with $\eta = \log\left(\frac{\theta}{1-\theta}\right)$,

$$h(x) = 1 \ , \quad u(x) = x \ , \quad \text{and} \quad g(\eta) = \frac{1}{1+\exp(\eta)} \ .$$

(Note that η contains sufficient information for θ, i.e., η is sufficient statistics for θ)

EM for Exponential Family

Given a complete data $s = (x, z)$ modeled by a distribution of exponential family, we write the expected complete-data log-likelihood:

$$\begin{aligned}\mathcal{Q}(\theta; \theta^{(t)}) &:= \mathbb{E}_{z|x,\theta^{(k)}}[\log p(s \mid \theta)] \\ &= \mathbb{E}_{z|x,\theta^{(k)}}[\eta(\theta)^\top u(s)] + \mathbb{E}_{z|x,\theta^{(k)}}[\log(h(s))] + \log g(\eta(\theta)) \\ &= \eta(\theta)^\top \mathbb{E}_{z|x,\theta^{(k)}}[u(s)] + \mathbb{E}_{z|x,\theta^{(k)}}[\log(h(s))] + \log g(\eta(\theta)) \,.\end{aligned}$$

Hence, the EM algorithm is given as:

- E-step: $u^{(k+1)} = \mathbb{E}_{z|x,\theta^{(k)}}[u(s)]$.
- M-step: $\theta^{(k+1)} = \arg\max_\theta \left[\eta(\theta)^\top u^{(k+1)} + \log g(\eta(\theta))\right]$.