ELBO

2024年10月28日 星期一 11:07

Expectation Maximization Intuition

EM is the standard approach to optimize this kind of problem,
Usually people say EM as a generalization of k-means, but here we are looking at a general formulation

Intuitively, EM convert the hard unsupervised problem into a simple supervised problem
We make an surrogate distribution to supervised optimized on, and use the optimized parameter to derive such surrogate distribution

M-step: Let's pretend we also observe the z distribution (then take expectation)

$$\max_\theta \mathbb{E}_{q(z|x)}[\log p(x, z|\theta)] \Leftarrow$$ Supervised MLE, just with one more z, do conditioned

E-step: We don't actually observe q, let's estimate it

$$q(z|x) = p(z|x, \theta) \Leftarrow$$ Posterior distribution of z given observed x and current θ parameter

Iterative process — Better q → θ, Better θ → q, End to end optimization

This is a very populate schema in ML: for instance in GAN

ELBO from Jensen's Inequality

Optimize a lowerbound

- For any distribution $q(z|x)$, define expected complete log likelihood:

$$\mathbb{E}_q[\ell_c(\theta; x, z)] = \sum_z q(z|x) \log p(x, z|\theta)$$ ← If we know q, we need to check all marginals that q gives from z|x, but no q now, could do anything

  - A deterministic function of θ
  - Inherit the factorizability of $\ell_c(\theta; x, z)$
- Use this as the surrogate objective — Difference objective
- Does maximizing this surrogate yield a maximizer of the likelihood?
  - We can show that:

$$\underbrace{\ell(\theta; x)}_{\text{Unsupervised loss}} \geq \underbrace{\mathbb{E}_q[\ell_c(\theta; x, z)] + H(q)}_{\text{Lower bound of such loss}}$$ (EM maximize the lower bound)

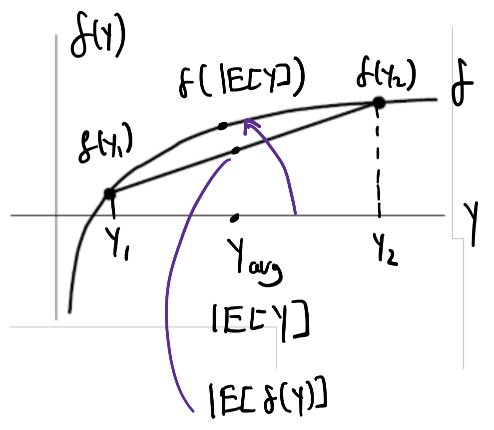
Jensen's Inequality

$$\mathbb{E}_{p(y)}[f(y)] \geq f(\mathbb{E}_{p(y)}[y])$$

If f is convex function, then the expectation can be a upperbound

Then if f is concave $\mathbb{E}_{p(y)}[f(y)] \leq f(\mathbb{E}_{p(y)}[y])$
Put function on avg — Put avg on function

Thinking about convexity from a probabilistic method!!!



Proof ELBO

$$L(\theta; x) = \log p(x|\theta) = \log \sum_z \overset{\text{marginal}}{p(x, z|\theta)}$$
$$= \log \sum_z p(z|x) \frac{p(x,z|\theta)}{p(z|x)}$$ ← multiply by $p(z|x)$
$$= \underset{\uparrow \text{ Concave function (the relationship is completely flipped)}}{\log} \mathbb{E}_{q(z|x)}\left[\frac{p(x,z|\theta)}{p(z|x)}\right]$$

Don't know what z is, we need to sum up all the z for finding out correct θ to match x, althrough z

$$\geq \mathbb{E}_{q(z|x)}\left[\log \frac{p(x,z|\theta)}{p(z|x)}\right]$$ Evidence of Lower Bound (ELBO)

Math is very simple

$$= \mathbb{E}_{q(z|x)}[\log p(x, z|\theta) - \log p(z|x)]$$
$$= \mathbb{E}_{q(z|x)}[\log p(x, z|\theta)] - \mathbb{E}_{q(z|x)}[\log p(z|x)]$$
$$\mathbb{E}_q[\ell_c(\theta; x, z)] + H(q)$$

(Bottom line, would not perform worst than ELBO)

$$L_c(\theta; x) \geq \mathbb{E}[L_c(\theta; x, z)] + H(q)$$ ← (We get better result if we max original, but the least we can do is also bounded)

Loss of θ to match x is higher than the expectation of loss for θ to match both x and q where q is from q distribution + the entropy of the q distribution
↑ Use of surrogate objective

ML is scalable because we can introduce a lot of approximations
Different philosophy of different discipline
— We care about the science and engineering, must have application ← Optimizable

Again: $\underset{\uparrow \text{ Loss based on } x}{L(\theta; x)} \geq \mathbb{E}_q[\underset{\uparrow \text{ Loss base on } q}{L_c(\theta; x, z)}] + H(q) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x, z|\theta)}{q(z|x)}\right]$ ELBO

Used in M-step — q distribution that we want to get

ELBO from KL

Used in E-step: It can be shown that original loss is decomposed to marginal, surrogate, and posterior

$$L(\theta; x) = \mathbb{E}_{q(z|x)}\left[\log \frac{p(x, z|\theta)}{q(z|x)}\right] + KL(q(z|x) \,||\, p(z|x, \theta))$$

We can show that this is a ELBO as well because KL is always ≥ 0

RHS ≥ LHS

In E-step, we want to set the q distribution to the posterior distribution

$$\underset{\uparrow \text{ original}}{\ell(\theta; x)} = \mathbb{E}_{q(z|x)}\underset{\uparrow \text{ ELBO}}{\left[\log \frac{p(x, z|\theta)}{q(z|x)}\right]} + \underset{\uparrow \text{ KL between q and Posterior}}{\mathrm{KL}(q(z|x) \,||\, p(z|x, \theta))}$$

Does not depend on q, but given whatever q, this holds,
this is sort of a decomposition process

L = ELBO Optimization + KL(True $\overset{\text{hidden}}{\text{dist}}$ || current estimated)

optimize $L(\theta; x)$ (max) depend on $\underset{\uparrow}{\max} \mathbb{E}_q\left[\log \frac{p(x,z|\theta)}{q(z|x)}\right]$

Ideally KL is zero
$L(\theta; x)$ not dependent on KL,

so $\underset{\text{max}}{L(\theta; x)} - \underset{\text{min}}{KL} = \underset{\text{max}}{\mathbb{E}}\left[\log \frac{p(x,z|\theta)}{q(z|x)}\right]$