6.867 Machine Learning — Fall 2019
Recitation Note: Variational Learning

Huitao Shen
huitao@mit.edu

Nov 8, 2019

1 Generative Models

	Discriminative Model	Generative Model
Data	$(X^{(1)}, Y^{(1)}), \ldots, (X^{(n)}, Y^{(n)}) \sim (X, Y)$	$X^{(1)}, \ldots, X^{(n)} \sim X$
Model	$p(Y\|X)$	$p(X, Z) = p(X\|Z)p(Z)$
Goal	$p(Y\|X)$ for training and classification	$p(X)$ for training, $p(Z\|X)$ for inference

In the first half of this course, we focused on discriminative models. Mathematically, discriminative models are $p(Y|X)$, where X is the feature and Y is the label. In order to train the model, we maximize the conditional likelihood $\prod_i p(Y^{(i)}|X^{(i)})$ on finite samples independently drawn from the data distribution (X, Y).

In the following weeks, we will turn to generative models. Mathematically, generative models are $p(X, Z)$, where Z can be label or latent variable. Instead of modeling the joint likelihood $p(X, Z)$, generative models are normally modeled as $p(Z; \theta)$ and $p(X|Z; \theta)$, where θ is the model parameter. Note that θ is not a random variable and in the following we might omit θ to make our notations simple. It follows $p(X, Z) = p(X|Z)p(Z)$.

As a concrete example, consider Gaussian mixture model. $p(Z) \sim \text{Categorical}(\boldsymbol{\pi})$ and $p(X|Z_i) \sim \text{N}(\mu_i, \sigma_i^2)$. Here latent variable Z is the component of the mixture and the model parameter $\theta = \{\boldsymbol{\pi}, \mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2\}$, where K is the total number of components.

Our goal is to compute the following two probabilities:

- Marginal likelihood $p(X)$. The model is trained by maximizing this probability.
- Conditional probability $p(Z|X)$. The latent variable can be inferred from this probability after model training.

Both probabilities are typically difficult to compute. $p(Z|X) = p(X|Z)p(Z)/p(X)$ involves computing

$$p(X) = \sum_Z p(X, Z) = \sum_Z p(X|Z)p(Z). \tag{1}$$

Except for simple models like Gaussian mixture with small K, the sample space of Z is usually exponentially large such that the summation is difficult to perform numerically, or even continuous such that the summation becomes an integral that cannot be carried out analytically. In these cases, $p(X)$ and $p(X|Z)$ are called intractable. In the following, we introduce one method to deal with intractable likelihoods.

2 Expectation-Maximization

Consider the log marginal likelihood:

$$\ln p(X) = \ln \frac{p(X,Z)}{p(Z|X)} = \sum_Z q(Z) \ln \frac{p(X,Z)}{q(Z)} \frac{q(Z)}{p(Z|X)} \tag{2}$$

$$= \underbrace{\sum_Z q(Z) \ln \frac{p(X,Z)}{q(Z)}}_{\text{ELBO } L(q;\theta)} \underbrace{- \sum_Z q(Z) \ln \frac{p(Z|X)}{q(Z)}}_{\text{KL}(q(Z)||p(Z|X))}. \tag{3}$$

In (2), $q(Z)$ can be arbitrary distribution and can depend on X as $q(Z|X)$. We also use the fact that $\sum_Z q(Z) = 1$ and $q(Z)/q(Z) = 1$. Generally, it is possible to have a different $q^{(i)}(Z)$ for every data point in the marginal likelihood of the entire dataset: $\sum_i \ln p(X^{(i)})$. However, in some models such as variational autoencoder, there exists single $q(Z|X)$ called recognition network that is shared for all $X^{(i)}$.

In (3), the first term $L(q;\theta)$ is called the evidence lower bound (ELBO), and the reason for this name will become clear shortly. Here we explicitly include θ dependence because $p(X,Z)$ depends on θ. The second term is the KL divergence between $q(Z)$ and $p(Z|X)$. Since $\text{KL}(q(Z)||p(Z|X)) \geq 0$ and $\text{KL}(q(Z)||p(Z|X)) = 0$ iff $q(Z) = p(Z|X)$, it follows

$$\ln p(X;\theta) \geq L(q;\theta). \tag{4}$$

This inequality can also be proved directly using Jensen's inequality

$$\ln p(X) = \ln \sum_Z p(X,Z) \frac{q(Z)}{q(Z)} = \ln \mathbb{E}_{Z\sim q(Z)} \left[\frac{p(X,Z)}{q(Z)} \right] \geq \mathbb{E}_{Z\sim q(Z)} \left[\ln \frac{p(X,Z)}{q(Z)} \right] = L(q;\theta). \tag{5}$$

Expectation-Maximization (EM) is an algorithm that iteratively maximizes ELBO:

- E-step: $q^t = \text{argmax}_q \, L(q;\theta^t) = p(Z|X;\theta^t)$.
- M-step: $\theta^{t+1} = \text{argmax}_\theta \, L(q^t;\theta)$.

If both E-step and M-step are performed exactly, EM algorithm increases the marginal likelihood $\ln p(X;\theta)$ monotonically, because

$$\ln p(X;\theta^t) \underset{\text{E-step}}{=} L(q^t;\theta^t) \underset{\text{E-step}}{\leq} L(q^t;\theta^{t+1}) \underset{\text{ELBO}}{\leq} \ln p(X;\theta^{t+1}). \tag{6}$$

Note that ELBO itself can be further decomposed into an expectation and an entropy term (Exercise 9, Problem 4):

$$L(q;\theta) = \mathbb{E}_{Z\sim q(Z)} \left[\ln \frac{p(X,Z)}{q(Z)} \right] = H(q) + \mathbb{E}_{Z\sim q(Z)}[\ln p(X,Z)]. \tag{7}$$

Without the entropy term, E-step yields (ignore iteration superscript t):

$$q = \text{argmax}\, \mathbb{E}_{Z\sim q(Z)}[\ln p(X,Z)] = \delta(Z - Z'), \text{ where } Z' = \underset{Z}{\text{argmax}} \ln p(X,Z). \tag{8}$$

Therefore, entropy term can be viewed as regularization to avoid q collapsing into a single delta distribution.

3 Variational Expectation-Maximization

An important caveat is that in reality, E-step often cannot be carried out exactly, except for simple models such as Gaussian mixture with small K. The reasons are at least two-fold: (i) The reason why we need EM algorithm in the first place is that the optimal solution to q, i.e. $p(Z|X)$ is intractable; (ii) Even if $p(Z|X)$ is tractable, the optimization is generally not convex and there is no guarantee we find the maximum. To deal with the first problem, we restrict the possible space of q to Q. This leads to the variational Expectation-Maximization (vEM) algorithm:

- E-step: $q^t = \operatorname{argmax}_{q \in Q} L(q; \theta^t) = \operatorname{argmin}_{q \in Q} \mathrm{KL}(q(Z)||p(Z|X; \theta^t))$, because $\ln p(X)$ does not depend on q.
- M-step: $\theta^{t+1} = \operatorname{argmax}_{\theta} L(q^t; \theta)$.

Essentially, $q(Z)$ is an approximation to the conditional distribution $p(Z|X)$. In vEM, q is often parameterized by parameters ϕ: $q = q(Z|X; \phi)$ such that the E-step becomes

$$\phi^t = \operatorname*{argmin}_{\phi} \mathrm{KL}(q(Z; \phi)||p(Z|X; \theta^t)). \tag{9}$$

For vEM, there is no guarantee that the marginal likelihood increases in every iteration.