CS385 Machine Learning Homework 4

Zhou Litao 518030910407 F1803016

March 23, 2021, Spring Semester

Exercise 1 Please explain the concept of the entropy, the cross entropy, and the KL divergence.

Solution.

1. Entropy is defined as

$$\text{entropy}(p) = E_p[-\log p(X)] = \sum_x p(x)[-\log p(x)] \tag{1}$$

Entropy can be considered as a measure of uncertainty. It can also be interpreted as the minimum average length of the encoding of a particular probability distribution, where the unit of entropy is 'bit'.

2. Cross entropy is defined as

$$\text{CE}(P||Q) = \sum_x p(x)[-\log q(x)] \tag{2}$$

It can be interpreted as the average encoding length using the distribution of Q to encode P.

3. KL divergence is defined as

$$\begin{aligned} \text{KL}(p|q) &= E_p[-\log q(X)] - E_p[-\log p(X)] = E_p[\log \frac{p(X)}{q(X)}] \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &= \text{CE}(P||Q) - \text{entropy}(p) \end{aligned} \tag{3}$$

It measures the dissimilarity between two distributions.

□

Exercise 2 Please explain how to understand the discriminative model and the logistic regression as "learning from errors."

Solution. For discriminative model with softmax output layer, we can derive the gradient of the log likelihood as

$$\begin{aligned} \frac{\partial}{\partial \theta} \log p_\theta(y \mid X) &= \frac{\partial}{\partial \theta} f_\theta(X)^\top (Y - p) \quad \text{Learn from errors} \\ &= \frac{\partial}{\partial \theta} f_\theta(X)^\top \left(Y - \mathrm{E}_\theta(Y \mid X)\right) \end{aligned} \tag{4}$$

Here, $\frac{\partial}{\partial \theta} f_\theta(X)^\top$ is a term independent of the output, while $(Y - p)$ is the difference of the truth and the prediction value. To optimize the loss function and decrease the gradient, we should always learn from 'error' $(Y - p)$.

For logistic regression, if we take the gradient of the 0-1 loss function, we get that

$$l'(\beta) = \sum_{i=1}^n \left[y_i X_i - \frac{e^{X_i^\top \beta}}{1 + e^{X_i^\top \beta}} X_i \right] = \sum_{i=1}^n (y_i - p_i) X_i \tag{5}$$

It can be seen that the gradient is the product of X_i which is independent of the output, and the error term $y_i - p_i$. This observation shows that the logistic model is self consistent. □

Exercise 3 Please explain how to understand the descriptive model and the logistic regression as "learning from the dream."

Solution. For discriminative model with softmax output layer, we can derive the gradient of the log likelihood as

$$\begin{aligned} \frac{\partial}{\partial \theta} \log p_\theta(X) &= \frac{\partial}{\partial \theta} f_\theta(X) - \frac{\partial}{\partial \theta} \log Z(\theta) \\ &= \frac{\partial}{\partial \theta} f_\theta(X) - \mathrm{E}_\theta \left[\frac{\partial}{\partial \theta} f_\theta(X) \right] \end{aligned} \tag{6}$$

Here, $\frac{\partial}{\partial \theta} f_\theta(X)$ refers to the actual gradient of the data distribution, while $\mathrm{E}_\theta \left[\frac{\partial}{\partial \theta} f_\theta(X) \right]$ is the average gradient of our estimation ("dream"). We will learn based on the difference of the actual world and the "dream". □

Exercise 4 For descriptive model, how to compute the term of $E_\theta[\frac{\partial f(x)}{\partial \theta}]$ on Page 22? In other words, how to sample x from the distribution of $p_\theta(x)$.

Solution. We can use Langevin Dynamics to sample x given the disrtibution of p_θ. Langevin Dynamics simulates the Brownian motion in the natural world. The general rule for Langevin Dynamics is presented as follows.

$$X_{t+\Delta t} = X_t - \frac{1}{2} U'\left(X_t\right) \Delta t + \sqrt{\Delta t} \varepsilon_t \tag{7}$$

Here $X_t - U'\left(X_t\right) \Delta t / 2$ decreases the energy, and the Brownian motion $\sqrt{\Delta t} \varepsilon_t$ increases the entropy. The Langevin dynamics decreases the KL-divergence between the distribution of X_t and p_θ monotonically.

We can iterate on $X_{t+\delta t}$ using the Langevin Dynamics. The accumulated samples will approximate p_θ

For the descriptive model, starting from a random noise, the sampling process is given as follows.

$$X_{t+\Delta t} = X_t + \frac{1}{2} \frac{\partial}{\partial x} \log p_\theta\left(x\right) \Delta t + \sqrt{\Delta t} \varepsilon_t \tag{8}$$

□

Exercise 5 For generative model, how to sample h from the distribution of $p_\theta(h|x)$ on Page 23?

Solution. We can use Langevin Dynamics to sample x given the disrtibution of $p_\theta(h|x)$. The general principle of Langevin Dynamics have been given in Exercise 4.

For the generative model, in particular , we can sample h_i from $p_\theta\left(h_i \mid X_i\right)$ by Langevin dynamics

$$h_{t+\Delta t} = h_t + \frac{1}{2} \frac{\partial}{\partial h} \log p_\theta\left(h, X_i\right) \Delta t + \sqrt{\Delta t} \varepsilon_t \tag{9}$$

where $-\log p_\theta\left(h, X_i\right)$ plays the role of energy. The Langevin dynamics decreases the KL-divergence between the distribution of h_t and $p_\theta\left(h \mid X_i\right)$ monotonically.

□