Lecture FYS5429, April 16, 2024

FYS5429/9429 APRIL 16, 2024

$$p(x_i, h_j; \Theta) = \frac{f(x_i, h_j; \Theta)}{Z(\Theta)}$$

joint probability

$$Z(\Theta) = \int_x dx \int_h dh\, p(x, h; \Theta)$$

$$\left(\sum_{x_i=1}^{M} \sum_{h_j}^{N} p(x_i, h_j; \Theta) \right)$$

$x_i = \{0, 1\}$ $h_j = \{0, 1\}$

$2^M 2^N$ possible configs.

Marginal probability

$$p(x;\theta) = \sum_{h_j} p(x, h_j; \theta)$$

$$p(h;\theta) = \sum_{x_k'} p(x_k', h; \theta)$$

Conditional probability

$$p(x|h;\theta) = \frac{p(x,h;\theta)}{p(h;\theta)}$$

$$p(h|x;\theta) = \frac{p(x,h;\theta)}{p(x;\theta)}$$

Bayes' theorem

$$p(x|h;\theta) = \frac{P(h|x;\theta)\,p(x;\theta)}{\sum_{x_i} p(h|x=x_i;\theta)\,p(x_i;\theta)}$$

Likelihood

prior

posterior probability

From last week

$$p(x;\theta) = \frac{1}{Z(\theta)} \prod_{x_i \in X} f(x_i;\theta)$$

$$= \frac{1}{Z(\theta)} \prod_{x_i} \left[\sum_{h_j} f(x_i, h_j;\theta) \right]$$

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\arg\max} \log p(x;\theta)$$

$$\nabla_\theta \log p(x;\theta) = 0$$

$$= \nabla_\theta \left[\sum_{x_i'} \log f(x_i'; \theta) \right]$$

$$- \mathbb{E}\left[\nabla_\theta \log f(x_i'; \theta) \right]$$

$$\sum_{x_i'} p(x_i'; \theta) \nabla_\theta \log f(x_i'; \theta)$$

Binary-Binary RBM

$$f(x_i', h_j'; \theta) = -\left(\sum_{i=1}^{M} a_i' x_i' + \sum_{j=1}^{N} b_j h_j' + \sum_{ij'} x_i' W_{ij}' h_j' \right)$$

$$= -\left(a^T x + b^T h + x^T W h\right)$$

$$P(x, h; \theta) = \frac{e^{a^T x + b^T h + x^T W h}}{Z(\theta)}$$

$$P(x; \theta) = \frac{1}{Z(\theta)} e^{a^T x} \prod_{j=1}^{N} \left(1 + e^{b_j + x^T W_{*j}}\right)$$

$$P(h; \theta) = \frac{1}{Z(\theta)} e^{b^T h} \prod_{i=1}^{M} \left(1 + e^{a_i + W_{i*}^T h}\right)$$

$$p(h_j = 1 \mid x) = \frac{1}{1 + e^{-(b_j + x^T W_{*j})}}$$

$$p(h_j = 0 \mid x) = 1 - p(h_j = 1 \mid x)$$

$$p(x_i = 1 \mid h) = \frac{1}{1 + e^{-(a_i + W_{i*}^T h)}}$$

$$p(x_i = 0 \mid h) = 1 - p(x_i = 1 \mid h)$$

$$f(x_i, h_j; \theta) = e^{-E(x_i, h_j; \theta)}$$

Kullback-Leibler divergence

$$KL[p \| q] = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

could be the PDF which governs our data

$q(x)$ is an approximation to $p(x)$.

Discrete case

$$KL[p \| q] = \sum_{x} p(x) \log \frac{p(x)}{q(x)}$$

if $p(x) = q(x)$, then

$KL[p \| q] = 0$

whenever $p(x)$ is zero,

$\lim_{x \to 0^+} x \log x = 0$

How do we use this for VAE?

we have our marginal distribution

$q(x) = p(x; \theta)$

in the KL-divergence

$p(x)$ represent our data, $f(x)$

$$KL\left[f(x)\,||\,p(x;\theta)\right]$$

$$= \int_{-\infty}^{\infty} f(x)\log\left(\frac{f(x)}{p(x;\theta)}\right)$$

$$= \int_{-\infty}^{\infty} f(x)\log f(x)\,dx$$

when we optimize it vanishes

$$- \int_{-\infty}^{\infty} f(x)\log p(x;\theta)\,dx$$

independent of θ

$$= \mathbb{E}_{x\sim f(x)}\left[\log f(x)\right] - \left(\mathbb{E}_{x\sim f(x)}\left[\log p(x;\theta)\right]\right)$$

What do we want?

$$p(x;\theta) = \int p(x \mid h;\theta)\, p(h)\, dh$$

(use max likelihood to find $\hat{\theta}$)

In VAE

$$p(x \mid h;\theta) \sim N(x \mid g(h;\theta), \sigma^2 \cdot \mathbb{1})$$

can use gradient descent.
Output not necessarily a Gaussian.

To solve for $p(x;\theta)$ we must deal with

- How to define the latent variables - h -
- How to integrate over h (sum)

Direct approach:
sample over a large set of $h = \{h_1, h_2, \ldots, h_n\}$

$$p(x;\theta) \simeq \frac{1}{n} \sum_{i=1}^{n} p(x \mid h_i;\theta)$$

We introduce a new function ($p(x|h)$) $q(h|x)$ which can take a value of -x- and give us a distribution over -h- values that are likely to produce x.

Hopefully, the space of -h- values under q is much smaller than the space of all h's that are likely to belong to prior $p(h)$

$\mathbb{E}_{h \sim q}$ $p(x|h)$ and $p(x;\theta)$ is one of the cornerstones of Bayesian statistics.

Bring Back KL-divergence

$$KL\left[q(h) \,\|\, p(h|x)\right] =$$

arbitrary function of h and may not depend on x.

$$\mathbb{E}_{u \sim q}\left[\log q(u) - \log p(u|x)\right]$$

$$= \int du\, q(u) \log q(u)$$

$$- \int du\, q(u) \log p(u|x))$$

How do we get $p(x)$ and $p(x|u)$ into this equation?

Apply Bayes' theorem.

$$p(x|u) = \frac{p(u|x)\, p(x)}{p(u)}$$

$$KL[q(h) \| p(h|x)] =$$

$$\left(E_{h\sim q}[\log q(h) - \log p(x|h)\right.$$

$$\left.- \log p(h)] + \log p(x)\right)$$

(taken out θ-dependence)

$$\log p(x) - KL[q(h) \| p(h|x)]$$

$$= \boxed{E_{h\sim q}[\log p(x|h)] - KL[q(h) \| p(h)]}$$

stochastic grad descent

x is fixed