Gradient-based Inferance for Refining the Approximate Inference in Variational Autoencoder with Discrete Latent Variables

September 22, 2015

1 Main

Let us define the joint log-probability of x and h as

$$\begin{aligned}
\log p(x,h) &= \log p(h) + \log p(x|h) & (1)\\
&= \sum_{k=1}^{K} (h_k \log \mu_k^h + (1-h_k)\log(1-\mu_k^h)) + \sum_{d=1}^{D} (x_d \log \mu_d^x + (1-x_d)\log(1-\mu_d^x)), & (2)
\end{aligned}$$

where μ^h is a parameter for the prior distribution, and $\mu^x = f(h)$. Let us use Ψ to denote μ^h as well as the parameters of f.

The objective in this case is to maximize the log probability of the marginal probability of x:

$$\begin{aligned}
\log p(x) &= \log \sum_h p(x,h)\\
&= \log \sum_h \tilde{q}(h|x) \frac{p(x,h)}{\tilde{q}(h|x)}\\
&\geq \sum_h \tilde{q}(h|x) \log \frac{p(x,h)}{\tilde{q}(h|x)}\\
&= \sum_h \tilde{q}(h|x) \log p(x,h) + \mathscr{H}(\tilde{q})\\
&\approx \frac{1}{N} \sum_{h^n} \log p(x,h^n) - \log q(h^n), \qquad (3)
\end{aligned}$$

where $\mathscr{H}(q)$ is the entropy of the approximate posterior q.

1.1 E Step: Approximately Inferring $p(h|x)$

Approximately inferring $p(h|x)$ is equivalent to

$$\begin{aligned}&\arg\max_{q} \sum_{h} \tilde{q}(h|x) \log p(x,h) + \mathscr{H}(\tilde{q}) \\ =&\arg\max_{\mu_1^h,\ldots,\mu_K^h} \sum_{h} \tilde{q}(h|x) \log p(x,h) + \mathscr{H}(\tilde{q}),\end{aligned}$$

where μ_k^h's are from Eq. (1).

The issue here is that we need to *sample* h's from $\tilde{q}(h|x)$, which results in a high-variance, computationally-expensive estimate. Instead, here we approximate it such that

$$\arg\max_{\mu_1^h,\ldots,\mu_K^h} \log p(x, \mu^h) + \mathscr{H}(\tilde{q}), \tag{4}$$

which is equivalent to approximate $f(h)$ with $f(\mu^h)$. This is equivalent to maximizing the lowebound of Eq. (3) (Need to check further).

Because there is a chance that this optimization may be non-convex, we initialize the optimization from a point $\mu^{h,0}$ given by a parametric function $q_\theta(h|x)$. In order to make sure that the initial point is close to the optimum, later in the M step, we add the following auxiliary cost function:

$$C_q(\theta) = \sum_{k=1}^{K} q_{\theta,k}(h|x) \log \mu_k^{h,*} + (1 - q_{\theta,k}(h|x)) \log(1 - \mu_k^{h,*}), \tag{5}$$

where $\mu^{h,*}$ is the solution found by Eq. (4). Need to check if the order is correct between $\mu^{h,*}$ and $q_{\theta,}(h|x)$.

1.2 M Step: Estimating the Parameters Ψ and θ

Just to recap: Ψ is a set of the parameters of the generation network, and θ is that of the recognition network.

Since the approximate inference has been computed, it is rather straightforward based on Eq. (3):

$$\arg\max_{\Psi} \frac{1}{N} \sum_{h^n} \log p_\Psi(x|h^n) + \log p_\Psi(h^n), \tag{6}$$

where h^n is the sample from the Bernoulli distribution with the parameters $\mu_1^{h,*}, \mu_2^{h,*}, \ldots, \mu_K^{h,*}$ obtained from the E-step.

Along with Eq. (6), we minimize Eq. (5) together.