hw1

October 14, 2018

1 CS 236: Deep Generative Models

Ramon Iglesias

1.1 Problem 1

$$\begin{aligned}
\text{argmax}_\theta \mathbb{E}[\log p_\theta(y|x)] &= \text{argmin}_\theta \mathbb{E}_{\hat{p}(x)}[D_{KL}(\hat{p}(y|x)||p_\theta(y|x)] \\
&= \text{argmin}_\theta \mathbb{E}_{\hat{p}(x)}[\mathbb{E}_{\hat{p}(y|x)}[\log \hat{p}(y|x) - \log p_\theta(y|x)]] \\
&= \text{argmin}_\theta \mathbb{E}_{\hat{p}(x)}[\mathbb{E}_{\hat{p}(y|x)}[\log \hat{p}(y|x)]] - \mathbb{E}_{\hat{p}(x)}[\mathbb{E}_{\hat{p}(y|x)}[\log p_\theta(y|x)]] \\
&= \text{argmin}_\theta \mathbb{E}_{\hat{p}(x,y)}[\log \hat{p}(y|x)] - \mathbb{E}_{\hat{p}(x,y)}[\log p_\theta(y|x)] \\
&= \text{argmax}_\theta \mathbb{E}_{\hat{p}(x,y)}[\log p_\theta(y|x)] - \mathbb{E}_{\hat{p}(x,y)}[\log \hat{p}(y|x)] \\
&= \text{argmax}_\theta \mathbb{E}_{\hat{p}(x,y)}[\log p_\theta(y|x)]
\end{aligned}$$

1.2 Problem 2

For simplicity, we denote p_θ as p and p_γ as $\hat{p}$.

We begin with Bayes rule:

$$p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)}$$

Note that $p(x|y)$ can be written in canonical form as:

$$p(x|y) = \exp(\frac{-1}{2}x^T\Lambda x + \eta_y^T x + g_y)$$

where,

$$\begin{aligned}
\Lambda &= \Sigma^{-1} = (\sigma^2 I)^{-1} \\
\eta_y &= \Lambda \mu_y \\
g_y &= \frac{-1}{2}\mu_y^T \Lambda \mu_y - \log((2\pi)^{\frac{n}{2}}|\sigma^2 I|^{\frac{1}{2}})
\end{aligned}$$

Thus,

$$\begin{aligned}
p(x|y)p(y) &= \exp(\frac{-1}{2}x^T\Lambda x + \eta_y^T x + g_y)\pi_y \\
&= \exp(\frac{-1}{2}x^T\Lambda x + \eta_y^T x + g_y + \log(\pi_y)) \\
&= \exp(\frac{-1}{2}x^T\Lambda x)\exp(\eta_y^T x + g_y + \log(\pi_y))
\end{aligned}$$

Plugging it back to $p(y|x)$:

$$
\begin{aligned}
p(y|x) &= \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)} \\
&= \frac{\exp(\frac{-1}{2}x^T\Lambda x)\exp(\eta_y^T x + g_y + \log(\pi_y))}{\sum_k \exp(\frac{-1}{2}x^T\Lambda x)\exp(\eta_k^T x + g_k + \log(\pi_k))} \\
&= \frac{\exp(\eta_y^T x + g_y + \log(\pi_y))}{\sum_k \exp(\eta_k^T x + g_k + \log(\pi_k))}
\end{aligned}
$$

Finally, by setting

$$
\begin{aligned}
\eta_y &= w_y \\
g_k + \log(\pi_y) &= b_y
\end{aligned}
$$

we have

$$
\begin{aligned}
p(y|x) &= \frac{\exp(\eta_y^T x + g_y + \log(\pi_y))}{\sum_k \exp(\eta_k^T x + g_k + \log(\pi_k))} \\
&= \frac{\exp(w_y^T x + b_y)}{\sum_k \exp(w_k^T x + b_y)} \\
&= \hat{p}(y|x)
\end{aligned}
$$

1.3 Problem 3

Q1 It would need $\prod_i^n k_i - 1$ parameters.

Q2 Let $\mathcal{N} := \{1, 2, \ldots, n\}$ be the indeces of the topological sort. The chain rule factorization can be written as:

$$p(X_1, \ldots, X_n) = \prod_{i=1}^{n} p(x_i|\mathcal{X}_i)$$

where $\mathcal{X}_i = \{x_j : i - m \leq j < i, j \in \mathcal{N}\}$.

For a given node $p(x_i|\mathcal{X}_i)$, the number of parameters needed to represent its conditional probability is:

$$(k_i - 1) \prod_{j \in \mathcal{X}_i} k_j$$

thus, for the entire factorization the total number of parameters is

$$\sum_{i \in \mathcal{N}} (k_i - 1) \prod_{j \in \mathcal{X}_i} k_j$$

Q3 When the variables are conditionally independent, the sets $\{\mathcal{X}_i\}_i$ become empty sets, i.e. $\mathcal{X}_i = \emptyset$ and the total number of parameters is

$$\sum_{i \in \mathcal{N}} (k_i - 1)$$

1.4 Problem 4

Q1 The forward and backward models do indeed cover the same hypothesis space. To see this, consider the case where $n = 2$, and note that the factorization described is simply the chain rule in a forward or reverse order:

$$p_f(x_1, x_2) = p_f(x_1)p_f(x_2|x_1)p_r(x_1, x_2) = p_r(x_2)p_r(x_1|x_2)$$

Using the canonical form of the gaussians, $\mathcal{N}(x|\mu, \sigma)$:

$$p(x) = \exp(\Lambda x^2 + \eta x + g)$$

where,

$$\begin{aligned}\Lambda &= \Sigma^{-1} = (\sigma^2 I)^{-1}\\ \eta &= \frac{\mu}{\sigma^2}\\ g &= \frac{-1}{2\sigma^2}\mu^2 - \log((2\pi)^{\frac{1}{2}}\sigma)\end{aligned}$$

We must show that $p_f = p_r$:

$$\begin{gathered}p_f(x_1)p_f(x_2|x_1) = p_r(x_2)p_r(x_1|x_2)\\ \mathcal{N}(x_1|u_1(0), \sigma_1^2(0))\mathcal{N}(x_2|u_2(x_1), \sigma_2^2(x_1)) = \mathcal{N}(x_2|\hat{u}_2(0), \hat{\sigma}_2^2(0))\mathcal{N}(x_1|\hat{u}_1(x_2), \hat{\sigma}_1^2(x_2))\\ \exp(\Lambda_1 x_1^2 + \eta_1 x_1 + g_1)\exp(\Lambda_{2|1} x_2^2 + \eta_{2|1} x_2 + g_{2|1}) = \exp(\hat{\Lambda}_2 x_2^2 + \hat{\eta}_2 x_2 + \hat{g}_2)\exp(\hat{\Lambda}_{1|2} x_1^2 + \hat{\eta}_{1|2} x_2 + \hat{g}_{1|2})\\ \Lambda_1 x_1^2 + \eta_1 x_1 + g_1 + \Lambda_{2|1} x_2^2 + \eta_{2|1} x_2 + g_{2|1} = \hat{\Lambda}_2 x_2^2 + \hat{\eta}_2 x_2 + \hat{g}_2 + \hat{\Lambda}_{1|2} x_1^2 + \hat{\eta}_{1|2} x_2 + \hat{g}_{1|2}\end{gathered}$$

Given $\Lambda_1, \eta_1, g_1, \Lambda_{2|1}, \eta_{2|1}, g_{2|1}$ and assuming sufficiently powerful neural networks, it is always possible to find $\Lambda_2, \eta_2, g_2, \Lambda_{1|2}, \eta_{1|2}, g_{1|2}$ such that the last equation holds.

1.5 Problem 5

Q1

$$\begin{aligned}\mathbb{E}_{z\sim p(z)}[\frac{1}{k}\sum_k p(x|z)] &= \frac{1}{k}\sum_k \mathbb{E}_{z\sim p(z)}[p(x|z)]\\ &= \frac{1}{k}\sum_k \int_z p(x|z)p(z)dz\\ &= \frac{1}{k}\sum_k \int_z p(x,z)dz\\ &= \frac{1}{k}\sum_k p(x)\\ &= p(x)\end{aligned}$$

Q2 Since $\log(x)$ is concave, Jensen's inequality shows that

$$\mathbb{E}[\log(x)] \leq \log(\mathbb{E}[x])$$

thus

$$\begin{aligned}\mathbb{E}_{z\sim p(z)}[\log(\frac{1}{k}\sum_k p(x|z))] &\leq \log(\mathbb{E}_{z\sim p(z)}[\frac{1}{k}\sum_k p(x|z)])\\ &= \log(p(x))\end{aligned}$$

Thus, $\log A$ is *not* an unbiased estimator of $\log p(x)$.

1.6 Problem 6

Q1 The number 656 in binary is 1010010000. Which requires 10 digits. Thus, $n = 10$ should be sufficient to represnt all 657 characters.

Q2 In binary, we can represent up to 1024 numbers using 10 digits. Thus, increasing the character alphabet from 657 to 900 would not increase the number of parameters n.

Q3-Q5 Submitted separately.