CS 236 Autumn 2019/2020 Homework [number]

SUNet ID: anuj42
Name: Anuj Shetty
Collaborators: [Shoaib Mohammed, Shrey Verma]

By turning in this assignment, I agree by the Stanford honor code and declare that all of this is my own work.

Problem 1

We notice that by Bayes' rule we can simplify $\mathbb{E}_{\hat{p}(x)}\mathbb{E}_{\hat{p}(y|x)} = \mathbb{E}_{\hat{p}(x,y)}$

$$\mathbb{E}_{\hat{p}(x)}\mathbb{E}_{\hat{p}(y|x)}[\log \hat{p}(y|x) - \log p_\theta(y|x)] = \mathbb{E}_{\hat{p}(x,y)}[\log \hat{p}(y|x) - \log p_\theta(y|x)]$$

We see that the first term involves only $\hat{p}$, which is independent of θ. Therefore, we can ignore it when we take the arg min with respect to θ.

$$\begin{aligned}
\arg\min_{\theta\in\Theta} \hat{E}_p(x)[D_{KL}(\hat{p}(y|x) \parallel p_\theta(y|x))] &= \arg\min_{\theta\in\Theta} \mathbb{E}_{\hat{p}(x,y)}[-\log p_\theta(y|x)] \\
&= \arg\max_{\theta\in\Theta} \mathbb{E}_{\hat{p}(x,y)}[\log p_\theta(y|x)]
\end{aligned}$$

Problem 2

$$\begin{aligned}
p_\theta(y) &= \pi_y, \text{ where } X_k \quad y = 1, \quad \pi_y = 1 \\
p_\theta(x|y) &= \mathcal{N}(x|\mu_y, \sigma^2 I) \\
p_\theta(y|x) &= \frac{p_\theta(x|y)p_\theta(y)}{\sum_y p_\theta(x|y)p_\theta(y)} \\
&= \frac{\pi_y \mathcal{N}(x|\mu_y, \sigma^2 I)}{\sum_{i=1}^k \pi_i \mathcal{N}(x|\mu_i, \sigma^2 I)} \\
&= \frac{\pi_y \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_y)^\top(x-\mu_y)\right)}{\sum_{i=1}^k \pi_i \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu_i)^\top(x-\mu_i)\right)} \\
&= \frac{\exp\left(-\frac{x^\top x}{2\sigma^2}\right) \exp\left(x^\top(\frac{\mu_y}{\sigma^2}) - (\frac{\mu_y^\top \mu_y}{2\sigma^2} - \log \pi_y)\right)}{\sum_{i=1}^k \exp\left(-\frac{x^\top x}{2\sigma^2}\right) \exp\left(x^\top(\frac{\mu_i}{\sigma^2}) - (\frac{\mu_i^\top \mu_i}{2\sigma^2} - \log \pi_i)\right)}
\end{aligned}$$

Comparing to

$$p_\gamma(y|x) = \frac{\exp(x^\top w_y + b_y)}{\sum_{i=1}^k \exp(x^\top w_i + b_i)}$$

We see that for any choice of θ, there exists γ such that $p_\theta(y|x) = p_\gamma(y|x)$, given by $w_i = \frac{\mu_i}{\sigma^2}, b_i = -\frac{\mu_i^\top \mu_i}{2\sigma^2} - \log \pi_i, \forall i$.

Problem 3

(1) $(\prod_i k_i) - 1$

(2) The variables would need to be fully independent of each other. If that holds, then simply specifying each $\Pr(X_i)$ with the required $k_i - 1$ independent parameters, would be sufficient to specify the joint distribution.

(3) We consider the variables $X_1, \ldots X_m$ first, which have no independence relation, therefore all i values need to be specified to determine each $\Pr(X_i|X_1, ..., X_{i-1})$, and we subtract 1 parameter since all probabilities must sum to 1. Then we consider each X_l for $l > m$, which is specified by the m variables before it.
Number of parameters $= (\sum_{j=1}^m \prod_{i=1}^j (k_i - 1)) + (\sum_{i=m+1}^n \prod_{j=i-m}^i k_j) - 1$

Problem 4

Let

$$\begin{aligned}
p_f(x_1) &= \mathcal{N}(x_1|0,1) \\
p_f(x_2|x_1) &= \mathcal{N}(x_2|0,0.1) \quad \text{if } x_1 < 0 \\
&= \mathcal{N}(x_2|0,10) \quad \text{if } x_1 \geq 0 \\
p_f(x_2) &= \int_{x_1} p_f(x_2|x_1) p_f(x_1)\, dx_1 \\
&= \int_{-\inf}^{0} \mathcal{N}(x_2|0,0.1)\mathcal{N}(x_1|0,1)\, dx_1 + \int_{0}^{\inf} \mathcal{N}(x_2|0,10)\mathcal{N}(x_1|0,1)\, dx_1 \\
&= \frac{1}{2}(\mathcal{N}(x_2|0,1.1) + \mathcal{N}(x_2|1,11))
\end{aligned}$$

$$\text{If } p_f = p_r, \quad \implies p_f(x_2) = p_r(x_2)$$

$$\text{However} \quad p_r(x_2) = \mathcal{N}(x_2|\hat{\mu}_2, \hat{\sigma}_2^2)$$

We can clearly see that a combination of 2 different Gaussians cannot be equal to a single Gaussian, therefore we have found a choice of $\{\mu_i, \sigma_i\}$ such that no choice of $\{\hat{\mu}_i, \hat{\sigma}_i\}$ can make $p_f = p_r$, so they are not equally expressive.

Problem 5

(1)

$$\begin{aligned}
\mathbb{E}[A(z^{(1)} \dots z^{(k)})] &= \mathbb{E}[\frac{1}{k}\sum_{i=1}^{k} p(x|z^{(i)})] \\
&= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}_z^{(i)}[p(x|z^{(i)})] \\
&= \frac{1}{k}\sum_{i=1}^{k} \int_z^{(i)} [p(z^{(i)})p(x|z^{(i)})]dz^{(i)} \\
&= \frac{1}{k}\sum_{i=1}^{k} \int_z^{(i)} [p(x, z^{(i)})]dz^{(i)} \\
&= \frac{1}{k}\sum_{i=1}^{k} [p(x)] \\
&= p(x)
\end{aligned}$$

Therefore A is an unbiased estimator of $p(x)$.

(2) No $log(A)$ is not an unbiased estimator of $\log(p(x))$. By Jensen's inequality, f(mean of x) is greater than or equal to mean of f(x) for concave f(x).

$$\begin{aligned}
\mathbb{E}[\log(A(z^{(1)} \dots z^{(k)}))] &= \mathbb{E}[\log(\frac{1}{k}\sum_{i=1}^{k} p(x|z^{(i)}))] \\
&>= \frac{1}{k}\sum_{i=1}^{k} \mathbb{E}[\log(p(x|z^{(i)}))] \\
&= \frac{1}{k}\sum_{i=1}^{k} \int_z^{(i)} [p(z^{(i)})[\log(p(x|z^{(i)}))]]dz^{(i)} \\
&>= \frac{1}{k}\sum_{i=1}^{k} \log(p(x)) \\
&= \log(p(x))
\end{aligned}$$

Since these are not equal, log A is not an unbiased estimator of log p(x)

Problem 6

(1) $\log_2 50257 = 15.6$. Therefore we need 16 bits as the most efficient representation for a token with 50257 unique values.

(2) Changing the number of possible tokens wil increase the number of parameters in the fully connected layer. This is because the number of parameters in the fully connected layer is equal to the number of tokens multiplied by the number of hidden units. Therefore, increasing the number of tokens from 50257 to 60000 will increase the number of parameters from 768*50257 (= 38701056) to 768*60000 (= 46080000), an **increase of 7378944**. The number of parameters in the GPT-2 model doesn't change, and softmax has no trainable parameters.