CS 236 Autumn 2019/2020 Homework 1

SUNet ID: 06009508
Name: Brandon McKinzie
Collaborators: N/A

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

Problem 1: MLE and KL Divergence

$$D_{KL}(\hat{p}(y \mid x)||p_\theta(y \mid x)) = \mathbb{E}_{\hat{p}(y|x)}\left[\log \frac{\hat{p}(y \mid x)}{p_\theta(y \mid x)}\right] \tag{1}$$

$$= \mathbb{E}_{\hat{p}(y|x)}\left[\log \hat{p}(y \mid x)\right] - \mathbb{E}_{\hat{p}(y|x)}\left[\log p_\theta(y \mid x)\right] \tag{2}$$

$$\arg\min_{\theta\in\Theta} \mathbb{E}_{\hat{p}(x)}\left[D_{KL}(\hat{p}(y \mid x)||p_\theta(y \mid x))\right] = \arg\min_{\theta\in\Theta} \mathbb{E}_{\hat{p}(x)}\left[\mathbb{E}_{\hat{p}(y|x)}\left[\log \hat{p}(y \mid x)\right] - \mathbb{E}_{\hat{p}(y|x)}\left[\log p_\theta(y \mid x)\right]\right] \tag{3}$$

$$= \arg\min_{\theta\in\Theta} \mathbb{E}_{\hat{p}(y,x)}\left[\log \hat{p}(y \mid x)\right] - \mathbb{E}_{\hat{p}(y,x)}\left[\log p_\theta(y \mid x)\right] \tag{4}$$

$$= -\arg\min_{\theta\in\Theta} \mathbb{E}_{\hat{p}(y,x)}\left[\log p_\theta(y \mid x)\right] \tag{5}$$

$$= \arg\max_{\theta\in\Theta} \mathbb{E}_{\hat{p}(y,x)}\left[\log p_\theta(y \mid x)\right] \tag{6}$$

Problem 2: Logistic Regression and Naive Bayes

We are asked to show that $\forall\theta$, $\exists\gamma$ such that

$$p_\theta(y \mid x) = p_\gamma(y \mid x) \tag{7}$$

First, note that we can rewrite $p_\theta(y \mid x)$ using Bayes' rule and basic probability:

$$p_\theta(y \mid x) = \frac{p_\theta(x \mid y)p_\theta(y)}{\sum_{y'} p_\theta(x \mid y')p_\theta(y')} \tag{8}$$

Then, substituting in the formulas we were provided:

$$\begin{align} p_\theta(y \mid x) &= \frac{\mathcal{N}(x \mid \mu_y, \sigma^2 I)\pi_y}{\sum_{y'} \mathcal{N}(x \mid \mu_{y'}, \sigma^2 I)\pi_{y'}} \tag{9} \\ &= \frac{\exp\left(-(x-\mu_y)^2/(2\sigma^2)\right)\pi_y}{\sum_{y'} \exp\left(-(x-\mu_{y'})^2/(2\sigma^2)\right)\pi_{y'}} \tag{10} \end{align}$$

where I've used compact notation to denote $(x-\mu_y)^T(x-\mu_y)$ simply as $(x-\mu_y)^2$. In what follows, let $\alpha = -\frac{1}{2\sigma^2}$. I'll also be writing $x^T x$ as x^2, etc. We can expand this out as follows:

$$\begin{align} \frac{\exp\left(-(x-\mu_y)^2/(2\sigma^2)\right)\pi_y}{\sum_{y'} \exp\left(-(x-\mu_{y'})^2/(2\sigma^2)\right)\pi_{y'}} &= \frac{\pi_y \cancel{e^{\alpha x^2}} e^{\alpha\mu_y^2} e^{-\alpha 2x^T\mu_y}}{\sum_{y'} \pi_{y'} \cancel{e^{\alpha x^2}} e^{\alpha\mu_{y'}^2} e^{-\alpha 2x^T\mu_{y'}}} \tag{11} \\ &= \frac{\pi_y e^{\alpha\mu_y^2} e^{-\alpha 2x^T\mu_y}}{\sum_{y'} \pi_{y'} e^{\alpha\mu_{y'}^2} e^{-\alpha 2x^T\mu_{y'}}} \tag{12} \\ & \tag{13} \end{align}$$

Therefore, for any θ, if we define γ using:

$$\begin{align} b_y &:= \log(\pi_y e^{\alpha\mu_y^2}) = \log(\pi_y) + \alpha\mu_y^2 \tag{14} \\ w_y &:= -2\alpha\mu_y \tag{15} \end{align}$$

then we will satisfy $p_\theta(y \mid x) = p_\gamma(y \mid x)$.

Problem 3: Conditional Independence and Parameterization

1. The total number of independent parameters is $(\prod_{i=1}^{n} k_i)$ - 1.

2. Our network factorizes the joint distribution as

$$p(X_1, X_2, \ldots, X_n) = p(X_1)\left[\prod_{i=2}^{m} p(X_i \mid X_{i-1}, \ldots, X_1)\right]\left[\prod_{i=m+1}^{n} p(X_i \mid X_{i-1}, \ldots, X_{i-m})\right] \tag{16}$$

   - $p(X_1)$ requires $k_1 - 1$ parameters.
   - For $1 < i \leq m$, $p(X_i \mid X_{i-1}, \ldots, X_1)$ requires $(\prod_{j=1}^{i} k_j) - 1$ parameters.
   - For $i > m$, $p(X_i \mid X_{i-1}, \ldots, X_{i-m})$ requires $(\prod_{j=i-m}^{i} k_j) - 1$ parameters.

   In total, the number of parameters is thus

$$(k_1 - 1) + \sum_{i=2}^{m}\left[(\prod_{j=1}^{i} k_j) - 1\right] + \sum_{i=m+1}^{n}\left[(\prod_{j=i-m}^{i} k_j) - 1\right] \tag{17}$$

3. To represent a Bayesian network over $X_1, \ldots, X_n$ with $\sum_{i=1}^{n}(k_i - 1)$ parameters, you'd need to impose $X_i \perp X_j \; \forall i, j \neq i$.

Problem 4: Autoregressive Models

Given any choice of $\{\mu_i, \sigma_i\}_{i=1}^n$, does there always exist a choice of $\{\hat{\mu}_i, \hat{\sigma}_i\}_{i=1}^n$ such that $p_f = p_r$?

As suggested by the hint, consider the case where $n = 2$. Then we have

$$p_f(x_1, x_2) = \mathcal{N}(x_1 \mid \mu_1, \sigma_1^2)\mathcal{N}(x_2 \mid \mu_2(x_1), \sigma_2^2(x_1)) \tag{18}$$
$$p_r(x_1, x_2) = \mathcal{N}(x_1 \mid \hat{\mu}_1(x_2), \hat{\sigma}_1^2(x_2))\mathcal{N}(x_2 \mid \hat{\mu}_2, \hat{\sigma}_2^2) \tag{19}$$

which reveals the answer that *no, these models do not cover the same hypothesis space of distributions.*

For example, if $\mu_1 = 0$ and σ_1^2 is sufficiently close to zero, then for negligibly small ϵ:

$$p_f(x_1, x_2) \approx \begin{cases} 0 & |x_1| < \epsilon \\ \mathcal{N}(x_2 \mid \mu_2(x_1), \sigma_2^2(x_1)) & |x_1| > \epsilon \end{cases} \tag{20}$$

The only way to ensure $p_r = p_f$ in this case is for $\mu_2(x_1)$ and $\sigma_2^2(x_1)$ to be constants (since $\hat{\mu}_2$ and $\hat{\sigma}_2^2$ are constants). Otherwise, $p_f(x_1, x_2)$ is essentially a gaussian over x_2 with moving mean and variance as a function of x_1, which p_r is not able to model.

Problem 5: Monte Carlo Integration

1. *Show that A is an unbiased estimator of $p(x)$.*

$$\begin{aligned}
\mathbb{E}_{p(z)}\left[A\right] &= \frac{1}{k}\sum_{i=1}^{k}\mathbb{E}_{p(z)}\left[p(x \mid z^{(i)})\right] &(21)\\
&= \frac{1}{k}\sum_{i=1}^{k} p(x) &(22)\\
&= p(x) &(23)
\end{aligned}$$

2. *Is $\log A$ and unbiased estimator of $\log p(x)$? Explain why or why not.* No, $\log A$ is not an unbiased estimator of $\log p(x)$, because $\mathbb{E}_{p(z)}\left[\log(f(z))\right] \neq \log \mathbb{E}_{p(z)}\left[f(z)\right]$[1].

[1] Note that if A were not a function of the $z^{(i)}$ variables, *then* (trivially) $\mathbb{E}_{p(z)}\left[\log(A)\right] = \log \mathbb{E}_{p(z)}\left[A\right] = \log A$

Problem 6: Programming Assignment

1. *Suppose we wish to find an efficient bit representation for the 50257 tokens. That is, every token is represented as $(a_1, \ldots, a_n)$, where $a_i \in \{0, 1\}, \forall i = 1, 2, \ldots, n$. What is the minimal n that we can use?*

   Since we are confined to use a fixed-length code of size n, we cannot exploit regularities/patterns in the data distribution. As such, the minimal n we can use is $\lceil \lg 50257 \rceil = 16$.

2. *If the number of possible tokens increases from 50257 to 60000, what is the increase in the number of parameters? Give an exact number and explain your answer.*

   The only layer that is impacted by this is the final fully-connected layer, which projects the transformer state to the output vocabulary space. The FC layer has a kernel with $768 \times V$ parameters (where V is vocabulary size)[2]. Therefore, increasing V from 50257 to 60000 results in a parameter increase of

   $$768 \times (60000 - 50257) \tag{24}$$

   which is equal to **7,482,624** additional parameters.

[2]GPT2 does not use a bias vector in their final projection. See line 171 of their model.py file.