CS236: Deep Generative Models Homework #1 Solution

(Fall 2021)

Shubham Shrivastava email: shubhams@stanford.edu

P1 Data distribution: $\hat{p}(x, y)$

inputs: $x \in \mathcal{X}$

outputs: $y \in \mathcal{Y}$

KL divergence is defined as:

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

We have to show that:

$$\arg\max_{\theta \in \Theta} E_{\hat{p}(x,y)}[\log p_\theta(y|x)] = \arg\min_{\theta \in \Theta} E_{\hat{p}(x)}[D_{KL}(\hat{p}(y|x) \| p_\theta(y|x))] \quad \text{(i)}$$

$\Rightarrow$ Let's expand the KL term:

$$D_{KL}(\hat{p}(y|x) \| p_\theta(y|x)) = E_{y \sim \hat{p}(y|x)}[\log \hat{p}(y|x) - \log p_\theta(y|x)] \quad \text{(ii)}$$

$$\Rightarrow \arg\min_{\theta \in \Theta} E_{\hat{p}(x)}[D_{KL}(\hat{p}(y|x) \| p_\theta(y|x))] = \arg\min_{\theta \in \Theta} E_{\hat{p}(x)}\left[E_{y \sim \hat{p}(y|x)}[\log \hat{p}(y|x) - \log p_\theta(y|x)]\right]$$

$$= \arg\min_{\theta \in \Theta} E_{\hat{p}(x)} E_{y \sim \hat{p}(y|x)} [\log \hat{p}(y|x)]$$

$$- \arg\min_{\theta \in \Theta} E_{\hat{p}(x)} E_{y \sim \hat{p}(y|x)} [\log p_\theta(y|x)]$$

$$E_{\hat{p}(x) \cdot \hat{p}(y|x)} = E_{\hat{p}(x,y)}$$

$$= \arg\min_{\theta \in \Theta} E_{\hat{p}(x,y)} [\log \hat{p}(y|x)]$$

$$- \arg\min_{\theta \in \Theta} E_{\hat{p}(x,y)} [\log p_\theta(y|x)] \quad \text{— (iii)}$$

The first term in above equation is independent of θ and hence does not contribute to the optimization problem – this term can be replaced by a constant 'c'. Also, utilizing the properties in equation (iv), we can re-write (iii) as (iv)

$$\max_\theta f(\theta) \cong \min_\theta [-f(\theta)] \quad \text{— (iv)}$$

So eq^n (iii) can be re-written as

$$\boxed{\arg\max_{\theta \in \Theta} E_{\hat{p}(x,y)} [\log p_\theta(y|x)]} \quad \text{— (v)}$$

P2 — The generative process is given as:

$$p_\theta(x|y) = \mathcal{N}(x|\mu_y, \sigma^2 I)$$

From Bayes rule:

$$p_\theta(y|x) = \frac{p_\theta(x|y)\, p_\theta(y)}{p_\theta(x)} \quad \text{— (i)}$$

where, $p_\theta(y)$ is given to be π_y ; $\sum_{y=1}^{k} \pi_y = 1$

, $p_\theta(x|y) = \mathcal{N}(x|\mu_y, \sigma^2 I)$

, $p_\theta(x)$ is a mixture of k gaussians

expanding the numerator in eq^n (i)

$$p_\theta(x|y)\, p_\theta(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_y)^T(x-\mu_y)}{2\sigma^2}\right) \cdot \pi_y \quad \text{— (ii)}$$

Also, $p_\theta(x)$ can be represented as:

$$p_\theta(x) = \sum_{i=1}^{k} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_i)^T(x-\mu_i)}{2\sigma^2}\right) \quad \text{— (iii)}$$

Substituting (ii) and (iii) in eqn (i)

$$p_\theta(y|x) = \frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_y)^T(x-\mu_y)}{2\sigma^2}\right)\pi_y}{\sum_{i=1}^{K} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_i)^T(x-\mu_i)}{2\sigma^2}\right)}$$

$$= \frac{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} \exp\left(-\frac{(x-\mu_y)^T(x-\mu_y)}{2\sigma^2}\right)\pi_y}{\cancel{\frac{1}{\sqrt{2\pi\sigma^2}}} \sum_{i=1}^{K} \exp\left(-\frac{(x-\mu_i)^T(x-\mu_i)}{2\sigma^2}\right)}$$

$$= \frac{\exp\left(\frac{-1}{2\sigma^2}\left(x^Tx - 2x^T\mu_y + \mu_y^T\mu_y\right)\right)\cdot\pi_y}{\sum_{i=1}^{K} \exp\left(\frac{-1}{2\sigma^2}\left(x^Tx - 2x^T\mu_i + \mu_i^T\mu_i\right)\right)}$$

$$= \frac{\pi_y \cdot \exp\left(\frac{-x^Tx}{2\sigma^2}\right) \Big/ \exp\left(\frac{1}{2\sigma^2}\left(-2x^T\mu_y + \mu_y^T\mu_y\right)\right)}{\sum_{i=1}^{K}\left[\exp\left(\frac{-x^Tx}{2\sigma^2}\right) \Big/ \exp\left(\frac{1}{2\sigma^2}\left(-2x^T\mu_i + \mu_i^T\mu_i\right)\right)\right]}$$

$$= \frac{\pi_y \cdot \exp\left(\frac{-x^T x}{2\sigma^2}\right) \Big/ \exp\left(\frac{1}{2\sigma^2}\left(-2x^T\mu_y + \mu_y^T\mu_y\right)\right)}{\sum_{i=1}^{K}\left[\exp\left(\frac{-x^T x}{2\sigma^2}\right) \Big/ \exp\left(\frac{1}{2\sigma^2}\left(-2x^T\mu_i + \mu_i^T\mu_i\right)\right)\right]}$$

$$= \frac{\pi_y \cdot \exp\left(\frac{-x^T x}{2\sigma^2}\right) \Big/ \exp\left(\frac{1}{2\sigma^2}\left(-2x^T\mu_y + \mu_y^T\mu_y\right)\right)}{K \cdot \exp\left(\frac{-x^T x}{2\sigma^2}\right) \Big/ \sum_{i=1}^{K}\exp\left(\frac{1}{2\sigma^2}\left(-2x^T\mu_i + \mu_i^T\mu_i\right)\right)}$$

$$= \frac{\exp\left(\frac{-1}{2\sigma^2}\left(2x^T\mu_y - \mu_y^T\mu_y\right)\right)\pi_y}{\sum_{j=1}^{K}\left[\exp\left(\frac{-1}{2\sigma^2}\left(2x^T\mu_i - \mu_i^T\mu_i\right)\right)+1\right]}$$

π_y can also be written as $\exp(\log_e \pi_y)$, and 1 can be rewritten as $\exp(\log_e \pi_i)$ where, $\pi_i = 1$

$$\Rightarrow p_\theta(y|x) = \frac{\exp\left(\frac{-1}{2\sigma^2}\left(2x^T\mu_y - \mu_y^T\mu_y + \log_e \pi_y\right)\right)}{\sum_{j=1}^{K}\left[\exp\left(\frac{-1}{2\sigma^2}\left(2x^T\mu_i - \mu_i^T\mu_i + \log_e \pi_i\right)\right)\right]}$$

$$= \frac{\exp\left(x^T\left(\frac{-2\mu_y}{2\sigma^2}\right) + \left(\frac{\mu_y^T\mu_y + \log_e \pi_y}{2\sigma^2}\right)\right)}{\sum_{j=1}^{K}\exp\left(x^T\left(\frac{-2\mu_i}{2\sigma^2}\right) + \left(\frac{\mu_i^T\mu_i + \log_e \pi_i}{2\sigma^2}\right)\right)} \quad \text{(iv)}$$

If we let $w_y = \frac{-2\mu_y}{2\sigma^2}$; $b_y = \frac{\mu_y^T\mu_y + \log_e \pi_y}{2\sigma^2}$

$$w_i = \frac{-2\mu_i}{2\sigma^2} \; ; \; b_i = \frac{\mu_i^T\mu_i + \log_e \pi_i}{2\sigma^2}$$

then equation (iv) becomes

$$p_\theta(y|x) = \frac{\exp\left(x^T w_y + b_y\right)}{\sum_{j=1}^{K}\exp\left(x^T w_i + b_i\right)} = p_y(y|x)$$

P3 (1) Without any conditional independence assumption, the joint distribution can be expressed using the chain rule:

i.e. $p(x_1, x_2, \cdots, x_n) = p(x_1)\, p(x_2|x_1) \cdots p(x_n|x_1 \cdots x_{n-1})$

Number of parameters → (k_1-1) $(k_2-1)k_1$ $(k_n-1)k_1 k_2 \cdots k_{n-1}$

So, total number of parameters:

$$(k_1-1) + k_1(k_2-1) + k_1k_2(k_3-1) + \cdots + k_1k_2\cdots k_{n-1}(k_n-1)$$

$$= (k_1-1) + \sum_{i=2}^{n} (k_i-1) \prod_{j=1}^{i-1} k_j$$

(2) If the joint distribution are assumed to be independent then they can be represented as

$$P(x_1, x_2, \cdots, x_n) = P(x_1)\, P(x_2) \cdots P(x_n)$$

Number of parameters needed to represent

$x_1 \to (k_1-1)$, $x_2 \to (k_2-1)$, $\cdots$, $x_n \to (k_n-1)$

So, total number of parameters: $\sum_{i=1}^{n} (k_i-1)$

(3) The joint distribution can be expressed as following using chain rule:

$$p(x_1, x_2, x_3, \cdots, x_n) = \underbrace{p(x_1, x_2, x_3, \cdots, x_m)}_{\text{no conditional independence assumption}} \cdot p(x_{m+1}, x_{m+2}, \cdots x_n \mid x_1, x_2, \cdots, x_m)$$

For the first term, number of parameters needed to specify the joint distribution is:

$$(K_1 - 1) + \sum_{i=2}^{m} (K_i - 1) \prod_{j=1}^{i-1} K_j$$

Let's expand the second term:

$$p(x_{m+1}, x_{m+2}, \cdots x_n \mid x_1, x_2, \cdots, x_m) = p(x_{m+1} \mid x_1, \cdots x_m) \, p(x_{m+2} \mid x_1, \cdots x_{m+1}) \cdots p(x_n \mid x_1, \cdots, x_{n-1})$$

Given the assumption that for $i > m$, the random variable x_i is conditional independent of all ancestors given the previous m ancestors, above eq^{n} can be re-written as:

$$p(x_{m+1}, x_{m+2}, \cdots x_n \mid x_1, x_2, \cdots, x_m) = p(x_{m+1} \mid x_1, \cdots x_m)$$
$$p(x_{m+2} \mid x_2, \cdots x_{m+1})$$
$$\vdots$$
$$p(x_n \mid x_{n-m} \cdots x_{n-1})$$

Number of parameters needed to specify the above distribution:

$$(k_{m+1}-1)\, k_1 k_2 \cdots k_m + (k_{m+2}-1)\, k_2 k_3 \cdots k_{m+1}$$
$$+ \cdots + (k_n - 1)\, k_{n-m}\, k_{n-m+1} \cdots k_{n-1}$$
$$= \sum_{i=m+1}^{n} (k_i - 1) \prod_{j=i-m}^{i-1} k_j$$

Hence, total number of parameters needed to specify the full joint:

$$(k_1 - 1) + \sum_{i=2}^{m} (k_i - 1) \prod_{j=1}^{i-1} k_j + \sum_{i=m+1}^{n} (k_i - 1) \prod_{j=i-m}^{i-1} k_j$$

P4 Let's consider the case where $n=2$

$$p_f(x_1, x_2) = p_f(x_1)\, p_f(x_2 \mid x_1)$$

where, $p_f(x_1) = \mathcal{N}(x_1 \mid \mu_1(0), \sigma_i^2(0))$

Since, $x_{<i} = \begin{cases} [x_1, \cdots x_{i-1}]^T & \text{if } i > 1 \\ 0 & \text{if } i < 1 \end{cases}$

It is given that $\mathbb{R}^0 = \{0\} \Rightarrow \mu_1(0) = c_1$ & $\sigma_1(0) = c_2$ for some constants c_1 & c_2

$$\Rightarrow p_f(x_1) = \mathcal{N}(x_1 \mid c_1, c_2^2)$$

$$p_f(x_2 \mid x_1) = \mathcal{N}(x_2 \mid \mu_2(x_1), \sigma_2^2(x_1))$$

Using Bayes rule: $p_f(x_1 \mid x_2) = \dfrac{p_f(x_2 \mid x_1)}{p_f(x_2)}$

where, $p_f(x_2) = \int_{x_1=-\infty}^{\infty} p_f(x_1)\, p_f(x_2 \mid x_1)\, dx_1$

$$= \int_{x_1=-\infty}^{\infty} \mathcal{N}(x_1 \mid c_1, c_2^2) \cdot \mathcal{N}(x_2 \mid \mu_2(x_1), \sigma_2^2(x_1))\, dx_1 \quad \text{(1)}$$

multi-modal gaussian

Let's look at the reverse function:

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

where, $p_r(x_2) = \mathcal{N}(x_2 | c_3, c_4^2)$ — (ii)

c_3, c_4: constants

It is clear from equations (i) and (ii) that $p_f(x_2)$ is a multivariate gaussian whereas $p_r(x_2)$ is a univariate gaussian. Hence, for some choice of $\{\mu_i, \sigma_i\}$, the neural network $\{\hat{\mu}_i, \hat{\sigma}_i\}$ cannot represent the same distribution.

PS ① A is given as:

$$A(z^{(1)}, \cdots, z^{(k)}) = \frac{1}{k} \sum_{j=1}^{k} p(x|z^{(j)}) \text{ , where } z^{(i)} \sim p(z)$$

Let's compute $E[A]$

$$E[A(z^{(1)}, \cdots, z^{(k)})] = \underbrace{\frac{1}{k} \sum_{j=1}^{k} E_{z^{(i)}}}_{p(z)} p(x|z^{(j)})$$

$$= p(z)\, p(x|z)$$

$$= \int_z p(x,z)\, dz$$

$$= p(x)$$

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

② From Jensen's inequality

$$E[\log y] \leq \log E[y]$$

$$\Rightarrow E[\log(A(z^{(i)}, \cdots, z^{(k)}))] \leq \log E[A(z^{(i)}, \cdots, z^{(k)}]$$

$\Rightarrow E[\log(A(z^{(i)}, \ldots, z^{(k)}))] \leq \log p(x)$ → from previous solution

An estimator $\hat{\theta}$ is an unbiased estimator of θ iff $E[\hat{\theta}] = \theta$. Here we have: $E[\log A] \leq \log p(x)$

Hence, $\log A$ is not an unbiased estimator of $\log(p(x))$

P6 ① 50257 tokens represented by n-bits

$(a_1, a_2, a_3, \cdots, a_n)$, where $a_i \in \{0, 1\}$, $\forall i = 1, 2, \cdots, n$

for minimal n, $2^n \geq 50,257$

$\Rightarrow n = 16$-bits

② Number of increased dimension: $60000 - 50257$

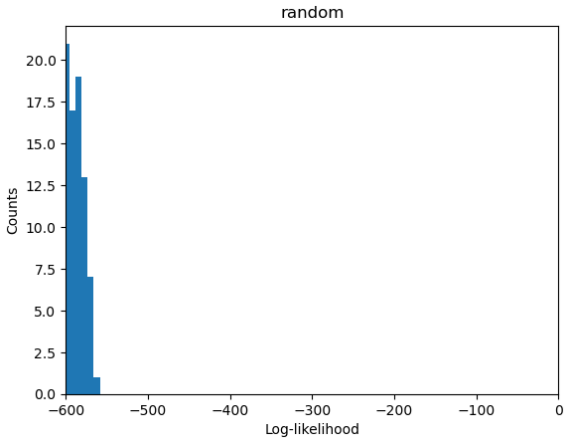
$= 9,743$

For fully-connected layers, number of additional parameters: $\underbrace{768 \times 9,743}_{\text{weights}} + \underbrace{9743}_{\text{Biases}} = 7,492,367$

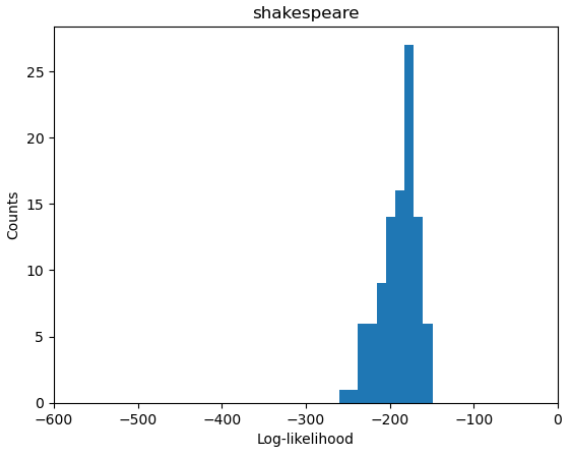
So, total increase in number of parameters:

7,492,367

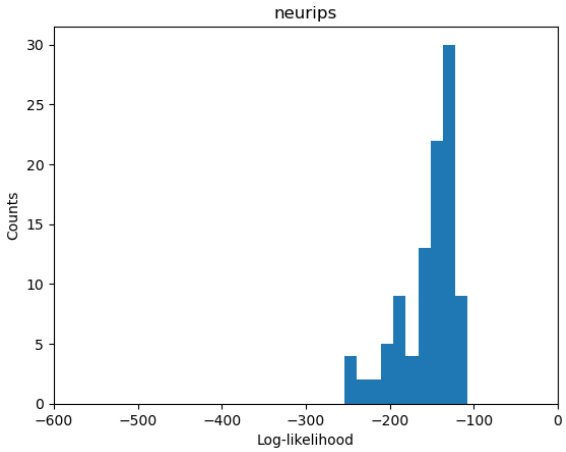
1 Question 6.4



(a) Histogram of the log-likelihoods of random strings



(b) Histogram of the log-likelihoods of string snippets from Shakespeare's work



(c) Histogram of the log-likelihoods of string snippets from NeurIPS 2015

Figure 1: Histogram of the log-likelihoods of strings.