Laboratorio 3

Autoencoders Variacionales

Profesor: Felipe Tobar
Auxiliares: Cristóbal Alcázar, Camilo Carvajal Reyes **Ayudante**: Joaquín Barceló

Fecha de entrega: 13 de octubre 2023

Instrucciones: El siguiente laboratorio consiste de una parte teórica y otra práctica, para la cual se adjuntará un código que deben completar con el formato solicitado. La ponderaciones son 50 % para cada parte. Deben entregar:

- Un reporte en formato PDF (extensión máxima de 4 páginas) con sus respuestas de la parte teórica. Puede ser a mano si lo desea, pero debe ser legible. NO es necesario entregar un infore de la parte práctica.
- Un archivo de extensión .ipynb donde contenga el código con sus respuestas a la parte práctica. Este debe seguir el formato de base (ser copia) del notebook entregado en material docente/repositorio.

Parte Teórica (50 %)

(**P1**) **Cota inferior de evidencia**
Definimos la divergencia de Kullback-Leibler entre dos medidas de probabilidad p y q como:

$$D_{KL}(p(x)\|q(x)) = \mathbb{E}_{x\sim p(\cdot)}\left(\log\left(\frac{p(x)}{q(x)}\right)\right)$$

a) (1.5 ptos) Demuestre la siguiente equivalencia,

$$D_{KL}(q(z|x)\|p(z|x)) = \log p(x) + \mathbb{E}_{z\sim q(\cdot|x)}(\log\left(\frac{q(z|x)}{p(x,z)}\right)),$$

donde $x \mapsto p(x|z)$ denota la medida de probabilidad condicional a z.

b) (1 pto) Demuestre que

$$\mathbb{E}_{z\sim q(\cdot|x)}(\log\left(\frac{q(z|x)}{p(x,z)}\right)) = D_{KL}(q(z|x)\|p(z)) - \mathbb{E}_{z\sim q(\cdot|x)}(\log p(x|z))$$

c) (1 pto) Usando las partes anteriores, escriba una **cota inferior** de la log-verosimilitud. Indicación: recuerde que la divergencia D_{KL} es siempre mayor o igual a cero.

(P2) VAE y reparametrización

Considere que la divergencia de Kullback-Leibler entre una distribución Gaussiana estándar $p(x) \sim \mathcal{N}(0, I)$ y una Gaussiana (no necesariamente estándar) $q(x) \sim \mathcal{N}(\mu, \sigma^2)$[1] tiene una forma cerrada dada por:

$$\begin{aligned} D_{KL}(q(z)\|p(z)) &= -\mathbb{E}_{z\sim q(\cdot)}\left(\log\left(\frac{p(z)}{q(z)}\right)\right) \\ &= -\frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right) . \end{aligned}$$

con $\mu = (\mu_1, \ldots, mu_J)$ y $\sigma = (\sigma_1, \ldots, \sigma_J)$ para J la dimensionalidad de la variable latente z. Además, podemos usar el truco de la reparametrización para reescribir, dada una función $f(\cdot)$, lo siguiente:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(\mu,\sigma^2)}\left(f(z)\right) = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left(f(\mu + \sigma\epsilon)\right) .$$

Además, podemos considerar una estimación de Monte-Carlo para la esperanza:

$$\mathbb{E}_{\epsilon\sim\mathcal{N}(0,1)}\left(f(\mu + \sigma\epsilon)\right) = \frac{1}{L}\sum_{l=1}^{L} f(\mu + \sigma\epsilon_l) ,$$

donde ϵ_l es una realización de $\mathcal{N}(0, 1)$ $\forall l = 1, \ldots, L$.

a) (1 pto) Considerando una formulación apropiada de la ELBO y usando lo anterior, de una aproximación de $\mathcal{L}_{\phi,\theta}(x)$ para un punto de dato x, en el caso donde $p_\theta(z) = \mathcal{N}(z; 0, I)$ y $q_\phi(z|x) = \mathcal{N}(z; \mu, \sigma^2)$.

Nota 1: μ y σ serán determinados posteriormente.
Nota 2: $p_\theta(z)$ no tiene parámetros, pero $p_\theta(x|z)$ si los tiene.

Un Autoencoder variacional corresponderá al caso anterior donde además computamos los elementos aprendibles a través de redes neuronales. Supondremos que nuestros datos son un subconjunto de los reales, i.e., $x \in \mathbb{R}^n$, por lo cual consideramos que la distribución $p_\theta(x|z)$ será una Gaussiana multivariada dado un $z \in \mathbb{R}^J$ fijo. Responda lo siguiente considerando la aproximación de la parte anterior.

b) (0.75 ptos) ¿Qué variables son modeladas por el encoder y el decoder respectivamente? Explique qué representan.

c) (Bonus) ¿Cómo generamos nuevos puntos de datos? Explique de qué modo logramos expresividad en los modelos pese a lo simple de las distribuciones.

d) (0.75 ptos) Explique cual es el problema de intentar usar descenso de gradiente (estocástico) para optimizar la ELBO antes del truco de la reparametrización.

[1] Como $\sigma = (\sigma_1, \ldots, \sigma_J)$ es un vector, $\mathcal{N}(\mu, \sigma^2)$ es un abuso de notación, donde la verdadera matriz de covariance es la matriz que tiene al vector σ^2 (el cuadrado de cada elemento) como diagonal.