Universidad de Chile
Facultad de Ciencias Físicas y Matemáticas
Departamento de Ingeniería Matemática

Laboratorio 3

Autoencoders Variacionales

Estudiante: Daniel Minaya
Profesor: Felipe Tobar
Auxiliares: Cristóbal Alcázar
Camilo Carvajal

Fecha de entrega: 13 de octubre

Cota inferior de evidecia

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) Demuestre la siguiente equivalencia,

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

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

Por el teorema de Bayes tenemos que $p(x,z) = p(z|x)p(x)$, luego

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

(b) Demuestre que

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

Por el teorema de Bayes tenemos que $p(x,z) = p(x|z)p(z)$, luego

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

(c) Usando las partes anteriores, escriba una cota inferior de la log-verosimilitud.

Usando que la divergencia KL es siempre no negativa, tenemos

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

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) $q(x) \sim \mathcal{N}(\mu, \sigma^2)$ tiene una forma cerrada dada por:

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

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, lo siguiente:

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

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

$$\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}(f(\mu + \sigma\varepsilon)) = \frac{1}{L}\sum_{\ell=1}^{L} f(\mu + \sigma\varepsilon_\ell),$$

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

(a) Considerando una formulación apropiada de la ELBO y usando lo anterior, entregue 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.

Tenemos que $\mathcal{L}_{\phi,\theta}(x) = \mathbb{E}_{z\sim q_\phi(\cdot|x)}(\log p_\theta(x|z)) - D_{KL}(q_\phi(z|x)||p_\theta(z))$. Reemplazando

$$\begin{aligned}
\mathcal{L}_{\phi,\theta}(x) &= \mathbb{E}_{z\sim\mathcal{N}(\mu,\sigma^2)}(\log p_\theta(x|z)) - D_{KL}(q_\phi(z|x)||p_\theta(z)) \\
&= \mathbb{E}_{\varepsilon\sim\mathcal{N}(0,1)}(\log p_\theta(x|\mu + \sigma\varepsilon)) + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right) \\
&\approx \frac{1}{L}\sum_{\ell=1}^{L}\log p_\theta(x|\mu + \sigma\varepsilon_\ell) + \frac{1}{2}\sum_{j=1}^{J}\left(1 + \log\sigma_j^2 - \mu_j^2 - \sigma_j^2\right),
\end{aligned}$$

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

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

El encoder modela la variable latente z. Su objetivo principal es codificar un dato $x \in \mathbb{R}^n$ en la variable $z \in \mathbb{R}^J$ con $J < n$. En otras palabras, el encoder transforma la entrada en un código de menor dimensión. Por otro lado, el decoder modela las variables de salida $\hat{x}$, con el propósito de generar una versión reconstruida a partir de un código z.

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

Para generar un dato lo que hacemos es samplear un punto z usando la distribución aprendida por el encoder durante el entrenamiento, luego pasamos z por el decoder para transformar z desde el espacio latente al espacio de datos, obteniendo así un dato generado $\hat{x}$.

La expresividad en un VAE se logra que el muestreo en el espacio latente es aleatorio, permitiendo generar una variedad de muestras distintas agregando mayor diversidad en las muestras generados. Además, recordemos que la dimensión del espacio latente es menor a la del espacio de los datos, por loque un VAE aprende a representar las características más significativas de los datos, lo que permite la generación de muestras más coherentes y expresivas.

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

El problema al usar descenso gradiente se encuentra en la parte de backpropagation. Debido a que z es sampleado de la distribución $\mathcal{N}(\mu, \sigma^2)$, entonces la loss function no es diferenciable con respecto a los parámetros μ y σ^2, por lo cual no es posible hacer la actualización. Para solucionar este problema se utiliza el truco de la reparametrización, lo que permite que el paso de x a z sea determinista, al dejar la aleatoriedad a una variable $\varepsilon \in \mathcal{N}(0, 1)$ independiente de los parámetros.