

## PUNTO UNO

7. Definición de algunos conceptos útiles más adelante

Regla de la credencia  $P(x, z) = P(x|z)P(z)$

Teorema de Bayes  $P(z|x) = \frac{P(x|z)P(z)}{P(x)}$

Divergencia Kullback-Leibler que proporciona la medida de la diferencia entre dos distribuciones de probabilidad.

$$D_{KL}(P||Q) = \int P(x) \log\left(\frac{P(x)}{Q(x)}\right) dx$$

Función esperanza  $E_x[F(x)] = \int x f(x) dx$

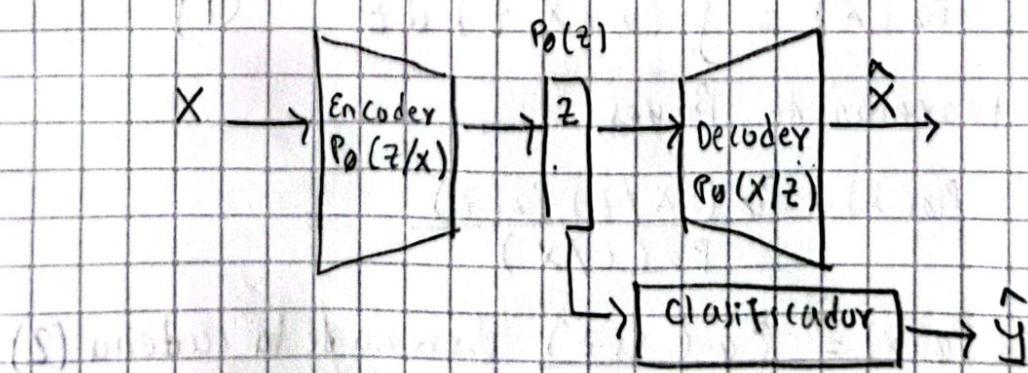
## 2. Autoencoder regularizado

Este autoencoder incorpora una restricción adicional en la representación aprendida por los datos. Esta restricción consiste en añadir un término de regularización al modelo para evitar sobreajuste. Por ejemplo:

Regularización  $L_1 \Rightarrow y_i l_i = \lambda \sum z_i$

controla la fuerza de regularización

### a. Modelo



Como se puede observar el autoencoder consta de:

\* Encoder  $z = f(w_e x + b_e)$  donde

datos caracteristicas  
↑ ↑  
 $N \times M$

$x$  es el conjunto de datos de entrada y  $x \in \mathbb{R}^M$   
 $w_e$  son los pesos del encoder y  $w_e \in \mathbb{R}^{M \times P}$ , nueva dimensión  
 $b_e$  es el bias del encoder y  $b_e \in \mathbb{R}^{N \times P}$   
 $f$  corresponde a la función de activación

\* Decoder  $\hat{x} = g(w_d z + b_d)$  donde

$z$  es la representación latente y  $z \in \mathbb{R}^P$   
 $w_d$  pesos del decodificador y  $w_d \in \mathbb{R}^{P \times M}$   
 $b_d$  bias del decodificador y  $b_d \in \mathbb{R}^{N \times M}$   
 $g$  es la función de activación  
 $\hat{x}$  es la entrada reconstruida y  $\hat{x} \in \mathbb{R}^{N \times M}$

\* Clasificador  $\hat{y} = h(w_c z + b_c)$  donde

$\hat{y}$  es la predicción y  $\hat{y} \in \mathbb{R}^{[0,1]^K}$  ( $K$  número de clases)  
 $w_c$  son los pesos del clasificador y  $w_c \in \mathbb{R}^{I \times N}$   
 $b_c$  son los bias del clasificador y  $b_c \in \mathbb{R}^{I \times P}$   
 $h$  es la función de activación

## b. Función de costos

En términos de probabilidad tenemos

$$P_0(x) = \int P_0(x, z) dz \quad (1)$$

y de acuerdo al teorema de Bayes

$$P_0(x) = \frac{P_0(x/z) P_0(z)}{P_0(z/x)}$$

$$P_0(x) = \underline{P_0(x/z)} \text{ por regla de la cadena (2)}$$

El cálculo de este término depende de una integral cosa su computación es difícil.  
Entonces se hace una aproximación

$$P_0(z/x) = q_{\phi}(z/x)$$

Se debe definir  $q_\phi$  de tal forma que sea lo más parecida a  $P_0(z)$  para obtener muestras similares a las originales. Luego, tenemos

$$P_0(x) = \frac{P_0(x, z)}{q_\phi(z|x)} \quad (3)$$

Subiendo que  $\int q_\phi(z|x)$  es uno porque se hace sobre el dominio de probabilidad

$$\log P_0(x) = \log P_0(x) \int q_\phi(z|x) dz$$

y teniendo en cuenta la definición de esperanza

$$E_{q_\phi(z|x)} [\log P_0(x)] = \int q_\phi(z|x) \log P_0(x) dz$$

Tenemos

$$\begin{aligned} \log P_0(x) &= E_{q_\phi(z|x)} [\log P_0(x)] \\ &= E_{q_\phi} \left[ \log \frac{P_0(x, z)}{P_0(z|x)} \right] \text{ Reemplazando (3)} \\ &= E_{q_\phi} \left[ \log \frac{P_0(x|z) P_0(z)}{P_0(z|x)} \right] \text{ Regla de la cadena} \\ &= E_{q_\phi} \left[ \log \frac{P_0(x|z)}{P_0(z|x)} + \log P_0(z) \right] \quad (4) \end{aligned}$$

Para inducir regularización L1 podemos fijar  $P_0(z) = \lambda \cdot 10^{-\lambda \sum z_i}$

$$\text{Entonces } \log P_0(z) = \log \lambda - \lambda \sum z_i$$

Entonces en (4)

$$\log P_0(x) = E_{q_\phi} \left[ \log \frac{P_0(x, z)}{P_0(z|x)} \right] + E_{q_\phi} \left[ \log \lambda - \lambda \sum z_i \right] \quad (5)$$

Del segundo término de (5)

$$\begin{aligned} E_{q_\phi} \left[ \log \lambda - \lambda \sum z_i \right] &= \int q_\phi(z) [\log \lambda - \lambda \sum z_i] dz = [\log \lambda - \lambda \sum z_i] \int q_\phi(z) dz \\ &= \log \lambda - \lambda \sum z_i \end{aligned}$$

la función de costos queda

$$\log P_\theta(x) = \mathbb{E}_{q_\phi} \left[ \log \frac{P_\theta(x, z)}{q_\phi(z|x)} \right] + \log \lambda - \lambda \sum_i z_i$$

Elbo (L)      Regularización

Diferencia entre valores originales  
y reconstruidos

### C. optimización

La optimización se hará sobre la maximización de L y minimización de  $\lambda \sum_i z_i$

$$\log P_\theta(x) \geq L$$

donde  $L = \mathbb{E}_{q_\phi(z|x)} \left[ \log \frac{P_\theta(x, z)}{q_\phi(z|x)} \right]$

Tomamos el gradiente  $\nabla_{\theta, \phi} L_{\theta, \phi}(x)$

El gradiente con respecto a  $\theta$  queda

$$\mathbb{E}_{q_\phi} \left[ \nabla_{\theta} \left[ \log P_\theta(x, z) - \log q_\phi(z|x) \right] \right]$$

$$= \mathbb{E}_{q_\phi} \nabla_{\theta} \log P_\theta(x, z) = \nabla_{\theta} \log P_\theta(x|z) =$$

El gradiente con respecto a  $\phi$  queda

$$\mathbb{E}_{q_\phi} \left[ \nabla_{\phi} \left[ \log P_\theta(x, z) - \log q_\phi(z|x) \right] \right]$$

Como la esperanza depende de  $\phi$ , es necesario implementar un cambio de variable

$$z = g(\phi, x, \epsilon) = \mu_\phi(x) + \sigma_\phi(x) \cdot \epsilon$$

$\hookrightarrow$  se obtiene una muestra de una distribución  $N(0, 1)$

$$\mathbb{E}_{q_\phi} [f(z)] = \mathbb{E}_\epsilon [f(g)] \text{ entonces } \nabla_{\phi} = \mathbb{E}_\epsilon \nabla_{\phi} [\log P_\theta - \log q_\phi]$$

### 3. Autoencoder Variacional

A diferencia del autoencoder regularizado, utiliza una función de pérdida que incluye la divergencia KL para regularizar el espacio latente.

#### a. Modelo

Tendrá los mismos elementos del autoencoder regularizado

#### b. Función de costos

Teniendo en cuenta el proceso inicial del autoencoder regularizado

$$\log P_0(x) = \mathbb{E}_{q\phi(z|x)} \left[ \log \frac{P_0(x, z)}{P_0(z|x)} \right]$$

Agregamos  $q\phi(z|x)$ , entonces

$$\begin{aligned} \log P_0(x) &= \mathbb{E}_{q\phi(z|x)} \left[ \log \frac{P_0(x, z)}{P_0(z|x)} \cdot \frac{q\phi}{q\phi} \right] \\ &= \mathbb{E}_{q\phi} \left[ \log \frac{P_0(x, z)}{q\phi} \right] + \mathbb{E}_{q\phi} \left[ \log \frac{q\phi}{P_0(z|x)} \right] \end{aligned}$$

Teniendo en cuenta la definición DKL, tenemos

$$DKL(q\phi || P_0(z|x)) = \int q\phi \log \frac{q\phi}{P_0(z|x)} = \mathbb{E}_{q\phi} \left[ \log \frac{q\phi}{P_0(z|x)} \right]$$

Entonces, la función de costos queda

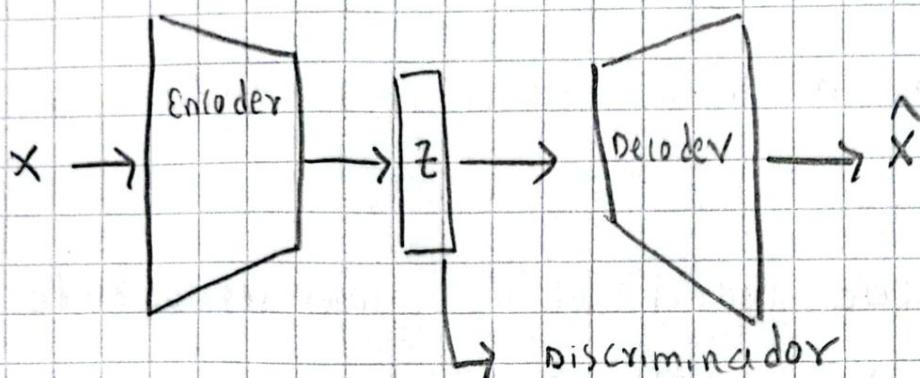
$$\log P_0(x) = \underbrace{\mathbb{E}_{q\phi} \left[ \log \frac{P_0(x, z)}{q\phi(z|x)} \right]}_{\text{Elbo LL}} + \underbrace{DKL(q\phi || P_0(z|x))}_{\text{Regularización}}$$

#### c. Optimización

Al igual que en el autoencoder regularizado, para optimizar la función de costos bastaría con optimizar el elbo con respecto a  $\phi$  y  $\theta$ .

## 4. GAN autoencoder

Este autoencoder busca aprovechar la capacidad de los autoencoders para aprender representaciones latentes de los datos y la capacidad de los GANs para generar datos realistas.



Como puede observarse el GAN consta de 3 elementos:

\* Codificador  $z = f(wx + bx)$  donde

$x$  es la entrada y  $x \in \mathbb{R}^{N \times M}$

$w$  son los pesos del decodificador y  $w \in \mathbb{R}^{M \times P}$

$b$  es el bias del decodificador y  $b \in \mathbb{R}^{N \times P}$

$f$  es la función de activación

\* Decodificador  $\hat{x} = g(w_d z + b_d)$  donde

$z$  es la representación latente y  $z \in \mathbb{R}^{N \times P}$

$w_d$  son los pesos del decodificador y  $w_d \in \mathbb{R}^{P \times M}$

$b_d$  es el bias del decodificador y  $b_d \in \mathbb{R}^{N \times M}$

$g$  es la función de activación

$\hat{x}$  es la entrada reconstruida

\* Discriminador  $D(x) = \sigma(w'_d x + b'_d)$  donde

$D(x)$  es la probabilidad de que  $x$  sea real

$w'_d$  son los pesos del discriminador

$b'_d$  son los bias del discriminador

$\sigma$  es la función sigmoidal

## b. Función de costos

Involucra dos funciones de pérdida (discriminador y decodificador)

$$V(D, G) = \mathbb{E} [\log(D(x))] + \mathbb{E} [\log(1 - D(G(z)))]$$

$x \sim P_{\text{data}}(x)$                            $z \sim P_z(z)$

mede la capacidad para  
distinguir entre datos reales  
y generados

mede la capacidad del ge-  
nerador para engañar el  
discriminador

## c. Optimización

El objetivo es maximizar la asignación de clase correcta  
y minimizar la capacidad para etiquetar datos generados

La razón por la cual se minimiza el generador  
en lugar de maximizarlo es porque  $\log(x < 1)$   
es negativo y  $D(G(z)) \approx 0\%$ ,  $G$  no  
engañaría al discriminador

$$D^* G = \max_D V(G, D) \quad \text{óptimo discriminador}$$

$$G^* = \min_G V(G, D^*) \quad \text{óptimo generador}$$

Lo anterior debe satisfacer  $P_G = P_{\text{data}}$

\* Teniendo en cuenta el random Nikodym theorem  
tenemos

$$\mathbb{E} [\log(1 - D(G(z)))] = \mathbb{E} [\log(1 - D(x))]$$

$z \sim P_z(z)$                            $x \sim P_{\text{data}}(x)$

Optimización para encontrar el óptimo discriminador

Derivamos la función de costos respecto a  $D(x)$

$$\frac{\partial U}{\partial D} = \frac{E_{x \sim P_{\text{data}}}}{D(x)} - \frac{E_{x \sim P_b(x)}}{1 - D(x)} = 0 \text{ maximizar}$$

$\Rightarrow$  operando la anterior expresión

$$D^*(x) = \frac{E_{x \sim P_{\text{data}}}}{E_{x \sim P_{\text{data}}} + E_{x \sim P_b(x)}}$$

Teniendo en cuenta la definición de esperanza

$$V(G, D) = \int P_{\text{data}}(x) \log D(x) dx + \int P_b(x) \log (1 - D(x)) dx$$

De modo que

$$D^*(x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_b(x)}$$

No conocemos  $P_{\text{data}}$ , así que no podemos usar directamente la expresión anterior

Nota: Si  $P_{\text{data}}$  tiene que ser igual a  $P_b$  entonces

$$D^*(x) = \frac{P_{\text{data}}(x)}{P_{\text{data}}(x) + P_{\text{data}}(x)} = \frac{1}{2}$$

$$D^*(x) = \frac{1}{2}$$

Optimización para encontrar el mejor generador

Si queremos lograr  $P_6 = P_{dato}$  entonces

$$D^*(x) = \frac{1}{2}$$

Podemos escribir  $V(G, D)$  como

$$\begin{aligned} V(G, D_6^*) &= \int P_{dato} \log \frac{1}{2} + P_6 \log \left( 1 - \frac{1}{2} \right) dx \\ &= -\log(2) \int P_6(x) dx - \log(2) \int P_{dato}(x) dx \\ &= -2\log(2) \int P_6(x) dx = -2\log(2) = -\log(4) \end{aligned}$$

El anterior valor sera un candidato para minimizar el generador y sucede cuando  $P_6 = P_{dato}$

Demostración

$$\begin{aligned} V(G, D) &= \int P_{dato} \log \left( \frac{P_{dato}}{P_{dato} + P_6} \right) + P_6 \log \left( 1 - \frac{P_{dato}}{P_{dato} + P_6} \right) dx \\ &\Rightarrow \frac{P_6}{P_{dato} + P_6} \end{aligned}$$

$\Rightarrow$

$$V(G, D) = \int P_{dato} \log \left( \frac{P_{dato}}{P_{dato} + P_6} \right) + P_6 \log \left( \frac{P_6}{P_{dato} + P_6} \right) dx$$

Con la finalidad de probar que  $-\log 4$  es el mínimo global añadimos y restamos  $\log(2)$

$$V(b, D) = \int (\log 2 - \log 2) P_{\text{data}} + P_{\text{data}} \log \left( \frac{P_{\text{data}}}{P_{\text{data}} + P_6} \right)$$

$$+ (\log 2 - \log 2) P_6 + P_6 \log \left( \frac{P_6}{P_{\text{data}} + P_6} \right) dx$$

$$= \int \cancel{\log 2} P_{\text{data}} - \cancel{\log 2} P_{\text{data}} + P_{\text{data}} \log \left( \frac{P_{\text{data}}}{P_{\text{data}} + P_6} \right)$$

$$+ \cancel{\log 2} P_6 - \cancel{\log 2} P_6 + P_6 \log \left( \frac{P_6}{P_{\text{data}} + P_6} \right) dx$$

$$= -\log(2) \int (P_6 + P_{\text{data}}) dx$$

$$+ \int P_{\text{data}} \left[ \log 2 + \log \frac{P_{\text{data}}}{P_{\text{data}} + P_6} \right] dx$$

$$+ \int P_6 \left[ \log 2 + \log \frac{P_6}{P_{\text{data}} + P_6} \right] dx$$

$$= -\log(2) [2]$$

$$+ \int P_{\text{data}} \log \frac{2P_{\text{data}}}{P_{\text{data}} + P_6} dx$$

$$+ \int P_6 \log \left( \frac{2P_6}{P_{\text{data}} + P_6} \right) dx$$

$$= -\log(4) + \int p_{data} \log \left( \frac{p_{data}}{(p_6 + p_{data})/2} \right) + p_6 \log \left( \frac{p_6}{(p_6 + p_{data})/2} \right) dx$$

Teniendo en cuenta la definición de DKL, tenemos

$$V(6, 0) = -\log 4 + D_{KL}(p_{data} \parallel (p_6 + p_{data})/2) \\ + D_{KL}(p_6 \parallel (p_6 + p_{data})/2)$$

Como  $D_{KL} > 0$  entonces  $-\log 4$  es el mínimo global

De esta manera se ha comprobado que  $V(D, 6)$  converge con un único valor a diferencia de Vay y sparse autoencoder

$$V(6, D^*) = -\log(4) \text{ se logra que } p_6 = p_{data}$$