CS 182 lecture 18: Latent Variable Models

Recap: $p(x) = \int p(x|z)\,p(z)\,dz$

$\theta \leftarrow \arg\max_\theta \frac{1}{N}\sum_i E_{z\sim p(z|x_i)}[\log p_\theta(x_i, z)] \Rightarrow \theta \leftarrow \arg\max_\theta \frac{1}{N}\sum_i \mathcal{L}_i(p, q_i)$

Variational approximation of $p(z|x_i)$
- ↳ $q_i(z) = N(\mu_i, \sigma_i)$. should approximate $p(z|x_i)$
- ↳ can bound $\log p(x_i)$.

$\log p(x_i) = \log \int_z p(x_i|z)\,p(z)$

$= \log \int_z p(x_i|z)\,p(z)\,\frac{q_i(z)}{q_i(z)}$

$= \log E_{z\sim q_i(z)}\left[\frac{p(x_i|z)\,p(z)}{q_i(z)}\right]$

$\geq E_{z\sim q_i(z)}\left[\log \frac{p(x_i|z)\,p(z)}{q_i(z)}\right]$

$= E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] - E_{z\sim q_i(z)}[\log q_i(z)]$

$= E_{z\sim q_i(z)}[\log p(x_i|z) + \log p(z)] + \mathcal{H}(q_i)$

↳ maximize this → maximizes $\log p(x_i)$

Jensen's inequality: $\log E[y] \geq E[\log y]$ (log is concave)

Entropy ("disorder"): $\mathcal{H}(p) = -E_{x\sim p(x)}[\log p(x)] = -\int_x p(x)\log p(x)\,dx$

KL-Divergence:
$D_{KL}(q\|p) = E_{x\sim q(x)}\left[\log\frac{q(x)}{p(x)}\right] = E_{x\sim q(x)}[\log q(x)] - E_{x\sim q(x)}[\log p(x)] = -(\mathcal{H}(q) + E_{x\sim q(x)}[\log p(x)])$
[how "different" two distributions are]

"Good" $q_i(z)$: $D_{KL}(q_i(z)\|p(z|x))$ should be low.
$= -\mathcal{L}_i(p, q_i) + \log p(x_i)$

$\log p(x_i) = D_{KL}(q_i(z)\|p(z|x_i)) + \mathcal{L}_i(p, q_i)$

min! ← causal! → max!
independent of q_i!

for each x_i (or B):
- → calculate $\nabla_\theta \mathcal{L}_i(p, q_i)$
  - → sample $z \sim q_i(z)$
  - → $\nabla_\theta \mathcal{L}_i(p, q_i) \approx \nabla_\theta \log p_\theta(x_i|z)$
- → $\theta \leftarrow \theta + \alpha\nabla_\theta \mathcal{L}_i(p, q_i)$
- → update q_i to maximize $\mathcal{L}_i(p, q_i)$. → $\phi \leftarrow \phi + \alpha\nabla_\phi \mathcal{L}$

use a second network:
$x \to q_\phi(z|x) = N(\mu_\phi(x), \sigma_\phi(x))$.

↘ $\log p(x_i) \geq$
$\left.\begin{array}{l} E_{z\sim q_\phi(z|x_i)}[\log p_\theta(x_i|z) + \log p(z)] \\ + \mathcal{H}(q_\phi(z|x_i)) \end{array}\right] \mathcal{L}_i$

$J(\phi) = E_{z\sim q_\phi(z|x_i)}[r(x_i, z)]$

$\nabla J(\phi) \approx \frac{1}{M}\sum_j \nabla_\phi \log q_\phi(z_j|x_i)\,r(x_i, z_j)$.

(policy gradient)

Substitute: $z = \mu_\phi(x) + \epsilon\sigma_\phi(x)$

$J(\phi) = E_{\epsilon\sim N(0,1)}[r(x_i, \mu_\phi(x_i) + \epsilon\sigma_\phi(x_i))]$

$\nabla_\phi J(\phi) \approx \frac{1}{M}\sum_j \nabla_\phi r(x_i, \mu_\phi(x_i) + \epsilon_j\sigma_\phi(x_i))$

single sample estimate of ϵ.

$\mathcal{L}_i \approx \log p_\theta(x_i|\mu_\phi(x_i) + \epsilon\sigma_\phi(x_i)) - D_{KL}(q_\phi(z|x_i)\|p(z))$ ← can be coded up.

"encoder"

$x_i \to \phi \rightrightarrows \mu_\phi(x_i),\ \sigma_\phi(x_i) \to \mu_\phi(x_i) + \epsilon\sigma_\phi(x_i) = z \to \theta \to p_\theta(x_i|z)$.

$\epsilon \sim N(0,1)$ – "noise"

VAE!

"decoder"

VAE objective: $\max_{\theta,\phi} \frac{1}{N}\sum_i \log p_\theta(x_i|\mu_\phi(x_i) + \epsilon\sigma_\phi(x_i)) - D_{KL}(q_\phi(z|x_i)\|p(z))$

encoder objective

"deviance"

Conditional models:

$\mathcal{L}_i = E_{z\sim q_\phi(z|x_i, y_i)}[\log p_\theta(y_i|x_i, z) + \log p(z|x_i)] + \mathcal{H}(q_\phi(z|x_i, y_i))$

$x_i, y_i \to \phi \rightrightarrows \mu_\phi(x_i, y_i),\ \sigma_\phi(x_i, y_i) \to \mu_\phi + \epsilon\sigma_\phi = z \to \theta \to p_\theta(y_i|x_i, z)$

$\epsilon \sim N(0,1)$, x_i

→ flexible w/ convolutions.

→ in practice: common for VAE/CVAE to ignore latent codes or not compress them at all.

↳ control D_{KL} carefully.

D_{KL} too low → $p_\theta(x|z) \to p(x)$. (blurry "average" image outputted)

D_{KL} too high → $q_\phi(z|x)$ very far from $p(z)$, "overfit", garbage output

use a β parameter to adjust regularizer strength.

should produce
- good reconstruction
- good samples
- perhaps be scheduled. increase β over time.

Invertible Model: $x = f(z)$ (deterministic)
(learn invertible mapping directly).

change of variables formula

$p(x) = p(z)\left|\det\left(\frac{df(z)}{dz}\right)\right|^{-1}$

change in local density due to f
(how small changes in $f(z)$ affect z)

$z = f^{-1}(x)$

Training Objective: $LL = \max_\theta \frac{1}{N}\sum_{i=1}^{N} \log p(x_i) \to \max_\theta \frac{1}{N}\sum_{i=1}^{N} \log p(f^{-1}(x_i)) - \log\left|\det\left(\frac{df(z)}{dz}\right)\right|$

Normalizing flow model: multiple layers of invertible transformations. (layers)

NICE:

$z_{1:d}$ → NN
$z_{d+1:n}$ → ⊕

$x_{1:d} = z_{1:d}$

$x_{d+1:n} = z_{d+1:n} + g_\theta(z_{1:d})$

e.g. $g_\theta(a) = \mathrm{ReLU}(Wa + b)$

$z_{1:d}$ ← NN ← $x_{1:d}$

$z_{d+1:n}$ ← ⊖ ← $x_{d+1:n}$

can scale further w/ Real-NVP.

$\begin{bmatrix} I & 0 \\ \frac{dg_\theta}{dz_{1:d}} & I \end{bmatrix}$

$\left|\det\left(\frac{df(z)}{dz}\right)\right| = 1!$ ↘ $p(x) = p(z)$.