Variational Auto-Encoder

Yingzi Bu

Department of Pharmaceutical Sciences
University of Michigan

Aug 21, 2023

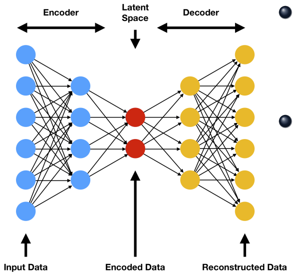
References

1. Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013). Cited by 28916
2. Kingma, Durk P., et al. "Semi-supervised learning with deep generative models." Advances in neural information processing systems 27 (2014). Cited by 3102
3. Gomez-Bombarelli, Rafael, et al. "Automatic chemical design using a data-driven continuous representation of molecules." ACS central science 4.2 (2018): 268-276. Cited by 2569
4. Zhavoronkov, Alex, et al. "Deep learning enables rapid identification of potent DDR1 kinase inhibitors." Nature biotechnology 37.9 (2019): 1038-1040. Cited by 769

Probability is all you need

$$
\begin{aligned}
&p(\text{The Sun will rise tomorrow}) && \approx 1 \\
&p(\text{cure cancer} \mid \text{compounds}) && = ??? \\
&p(\text{compounds} \mid \text{cure cancer}) && = ??? \\
&p(\text{approximation, interesting} \mid \text{Biologist, Chemist}) && \approx 1 \\
&p(\text{Talks in math} = \text{boring} \mid \text{Biologist, Chemist}) && > 0.5 \\
&p(\text{Math proof} \mid \text{you want to know ML}) && = 1
\end{aligned}
$$

Autoencoders



- Two sets:
  - the space of decoded messages $\mathscr{X} \subseteq \mathbb{R}^n$;
  - the space of encoded messages $\mathscr{Z} \subseteq \mathbb{R}^m$
- Two parametrized families of functions:
  - the encoder family: $E_\phi : \mathscr{X} \to \mathscr{Z}$, parametrized by ϕ
  - the decoder family: $D_\theta : \mathscr{Z} \to \mathscr{X}$, parametrized by θ

for $x \in \mathscr{X}$, latent variable $z = E_\phi(x) \in \mathscr{Z}$
for $z \in \mathscr{Z}$, decoded message $x' = D_\theta(z) \in \mathscr{X}$

Training Autoencoders

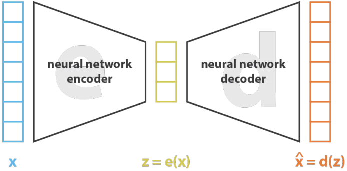
A reference probability distribution μ_{ref} over $\mathscr{X}$,
a function $d : \mathscr{X} \times \mathscr{X} \to [0, \infty)$ measures how much x' differs from x

$$L(\theta, \phi) := \mathbb{E}_{x \sim \mu_{\text{ref}}}[d(x, D_\theta(E_\phi(x))]$$

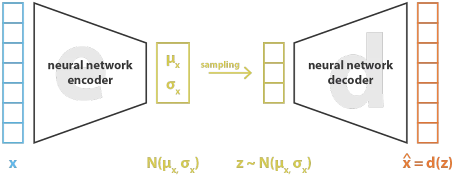
The least-squares optimization:

$$\min_{\theta,\phi} L(\theta, \phi), \text{ where } L(\theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} ||x_i - D_\theta(E_\phi(x_i))||_2^2$$

Variational Autoencoder (VAE)



$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$



$$\text{loss} = \| x - \hat{x} \|^2 + \mathrm{KL}[\, N(\mu_x, \sigma_x), N(0, I) \,] = \| x - d(z) \|^2 + \mathrm{KL}[\, N(\mu_x, \sigma_x), N(0, I) \,]$$

Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

Evidence lower bound (ELBO)

p^* true distribution of x, p_θ estimated distribution of x
Objective: $\max \mathbb{E}_{x \sim p^*(\cdot)}[\log p_\theta(x)]$
Alternative objective: $\max ELBO = \max_{\theta,\phi} \mathbb{E}_{z \sim q_\phi(\cdot|x)} log \frac{p_\theta(x,z)}{q_\phi(z|x)}$

$$
\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{x \sim p^*(\cdot)}[\log p_\theta(x)] \\
&= KL(q_\phi(z|x)||p_\theta(z|x)) + \mathbb{E}_{q_\phi(z|x)}[-\log q_\phi(z|x) + \log p_\theta(x,z)] \\
&\geq ELBO = -KL(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]
\end{aligned}
$$

$$
\tilde{\mathscr{L}}_{\text{VAE}}(x;\theta,\phi) = -KL(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x|z^{(l)})
$$

Loss function

$$\begin{aligned}
\min Loss(x;\theta,\phi) &= \min -\tilde{\mathscr{L}}_{\text{VAE}}(x;\theta,\phi) \\
&= \min KL(q_\phi(z|x)||p_\theta(z)) - \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)}) \\
&= \min \left[KL(\mathcal{N}(\mu,\mathbf{\Sigma})||\mathcal{N}(0,\boldsymbol{I})) - \frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x|z^{(l)}) \right] \\
&= \min \left[(-\frac{1}{2}\sum_{j=1}^{J}(1+\log\sigma_j^2-\mu_j^2-\sigma_j^2)) - \sum_{l=1}^{L}\log p_\theta(x|z^{(l)}) \right] \\
&= \min(\text{KLD} + \text{BCE / MSE})
\end{aligned}$$

```python
def loss_function(recon_x, x, mu, logvar):
    BCE = F.binary_cross_entropy(recon_x.cuda(),
                                 x.view(-1, 784).cuda(),
                                 reduction='sum')
    KLD = -0.5 * torch.sum(1 + logvar - mu.pow(2) - logvar.exp())
    return BCE + KLD
```

AE vs VAE generation, 100 epochs on MNIST data set

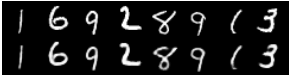


Figure 3: AE $x' \sim p_\theta(*|x)$

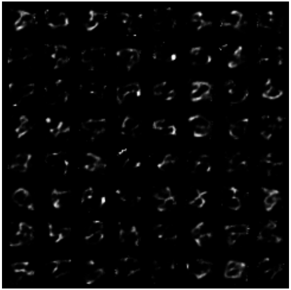


Figure 4: AE $x' \sim p_\theta(*|z)$,
$z \sim \mathcal{N}(0, \boldsymbol{I})$

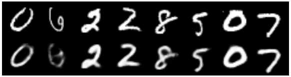


Figure 5: VAE $x' \sim p_\theta(*|x)$

Figure 6: VAE $x' \sim p_\theta(*|z)$,
$z \sim \mathcal{N}(0, \boldsymbol{I})$

ML question: AE latent space does not follow normal distribution anyway.

conditional generative models

What if I want compounds that have certain properties?

$$
\begin{aligned}
q_\phi(\boldsymbol{z}|\boldsymbol{x}, y) &= \mathcal{N}(\boldsymbol{\mu}_\phi(\boldsymbol{x}), \mathrm{diag}(\boldsymbol{\sigma}^2_\phi(\boldsymbol{x}))) \\
p_\theta(\boldsymbol{z}) &= \mathcal{N}(0, \boldsymbol{I}) \\
\boldsymbol{x} &\sim p_\theta(\cdot|\boldsymbol{z}, y)
\end{aligned}
$$

Question for audience: What will the architecture look like?

Math proof is left for your practice.

Stacked generative semi-supervised model

Only a subset of the observations have corresponding class labels.
How to make use of observations without labels?

Latent-feature discriminative model (M1):
VAE $p(z) = \mathcal{N}(z|0, I), p_\theta(x|z) = f(x; z, \theta)$

Generative semi-supervised model (M2)
VAE $p(y) = \mathrm{Cat}(y|\pi), p(z) = \mathcal{N}(z|0, I), p_\theta(x|y, z) = f(x; y, z, \theta)$

Stacked generative semi-supervised model (M1 + M2)
z_1 from M1, learn model M2 using embeddings from z_1 instead of x.
Two layers of stochastic variables:
$p_\theta(x, y, z_1, z_2) = p(y)p(z_2)p_\theta(z_1|y, z_2)p_\theta(x|z_1)$

M1: $p(\mathbf{z_1}|x) = \mathcal{N}(z|\mu, \mathrm{diag}(\sigma^{\mathbf{2}})), p(\mathbf{z_1}) = \mathcal{N}(z|0, I)$, decoder: $x \sim p_\theta(x|\mathbf{z_1})$
M2: $p(\mathbf{z_2}|\mathbf{z_1}, y) = \mathcal{N}(\mathbf{z_2}|\mu_{\mathbf{1}}, \mathrm{diag}(\sigma_{\mathbf{1}}^{\mathbf{2}})), p(\mathbf{z_2}) = \mathcal{N}(\mathbf{z_2}|0, I)$, decoder: $\mathbf{z_1} \sim p_\theta(\mathbf{z_1}|\mathbf{z_2}, y)$

Kingma, Durk P., et al. "Semi-supervised learning with deep generative models." Advances in neural information processing systems 27 (2014).

Stacked generative semi-supervised model: M1 loss

M1 loss:

$p(z|x) = \mathcal{N}(z|\mu, \text{diag}(\sigma^2)), p(z) = \mathcal{N}(z|0, I)$, decoder: $x \sim p_\theta(x|z)$

$\log p_\theta(x) \geq \mathbb{E}_{q_\theta(z|x)}[\log p_\theta(x|z) - KL(q_\phi(z|x)||p_\theta(z))] = -\mathscr{J}(x)$

Stacked generative semi-supervised model: M2 Loss

M2 loss:
$p(z_2|z_1, y) = \mathscr{N}(z_2|\mu_1, \text{diag}(\sigma_1^2)), p(z_2) = \mathscr{N}(z_2|0, I)$,
decoder: $z_1 \sim p_\theta(z_1|z_2, y)$, ($x$ means z_1, z means z_2) for eqs. below:

$$
\begin{aligned}
\log p_\theta(x, y) &\geq \mathbb{E}_{q_\phi(z|x,y)}[\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\phi(z|x, y)] \\
&= -\mathscr{L}(x, y) \text{ data with labels} \\
\log p_\theta(x) &\geq \mathbb{E}_{q_\phi(y,z|x)}[\log p_\theta(x|y, z) + \log p_\theta(y) + \log p(z) - \log q_\phi(y, z|x)] \\
&= \sum_y q_\phi(y|x)(-\mathscr{L}(x, y)) + \mathscr{H}(q_\phi(y|x)) = -\mathscr{U}(x) \text{ data no labels} \\
\mathscr{J} &= \sum_{(x,y)\sim\tilde{p}_l} \mathscr{L}(x, y) + \sum_{x\sim\tilde{p}_u} \mathscr{U}(x) \\
\mathscr{J}^\alpha &= \mathscr{J} + \alpha \cdot \mathbb{E}_{\tilde{p}_l(x,y)}[-\log q_\phi(y|x)]
\end{aligned}
$$

Derive ELBO

KL divergence: $KL(q||p) = \int q \log \frac{q}{p} = \mathbb{E}_q[\log \frac{q}{p}] \geq 0$, =0 holds iff $q = p$.

We want $p_\theta(x) \approx p^*(x)$, which is equivalent of $\min KL(p^*(x)||p_\theta(x))$

$$\begin{aligned}
\log p_\theta(x) &= \mathbb{E}_{x\sim p^*(\cdot)}[\log p_\theta(x)] = \mathbb{E}_{x\sim p^*(\cdot)}[\log \frac{p^*(x)}{p^*(x)} p_\theta(x)] \\
&= \mathbb{E}_{x\sim p^*(\cdot)} \log p^*(x) + \mathbb{E}_{x\sim p^*(\cdot)}[\log \frac{p_\theta(x)}{p^*(x)}] \\
&= -H(p^*) - KL(p^*(x)||p_\theta(x))
\end{aligned}$$

$\max \log p_\theta(x)$ is $\min KL(p^*(x)||p_\theta(x))$ as $H(p^*) = const.$

Then we need to find a way of calculating $\log p_\theta(x)$

Appendix

$$p_\theta(x) = \int p_\theta(x|z)p(z)dz = \int p_\theta(x,z)dz$$

$$= \int \frac{p_\theta(x,z)}{q_\phi(z|x)} q_\phi(z|x)dz = \mathbb{E}_{z\sim q_\phi(z|x)} \frac{p_\theta(x,z)}{q_\phi(z|x)}$$

$$\log p_\theta(x) = \log \mathbb{E}_{z\sim q_\phi(z|x)} \frac{p_\theta(x,z)}{q_\phi(z|x)} \geq \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

$$ELBO := \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right]$$

You can also see that:

$$ELBO = \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(z|x)p_\theta(x)}{q_\phi(z|x)}\right]$$

$$= \mathbb{E}_{z\sim q_\phi(z|x)} \left[\log \frac{p_\theta(z|x)}{q_\phi(z|x)}\right] + \mathbb{E}_{z\sim q_\phi(z|x)} [\log p_\theta(x)]$$

$$= -KL(q_\phi(z|x)||p_\theta(z|x)) + \log p_\theta(x) \leq \log p_\theta(x)$$

Appendix

$\min KL(p*(x)||p_\theta(x)) \to \max \log p_\theta(x) \to \max ELBO$. How to $\max ELBO$?

$$
\begin{aligned}
ELBO &= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}\right] \\
&= \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log \frac{p_\theta(z)}{q_\phi(z|x)}\right] + \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&= -KL(q_\phi(z|x)||p_\theta(z)) + \mathbb{E}_{z\sim q_\phi(z|x)}\left[\log p_\theta(x|z)\right] \\
&\approx -KL(q_\phi(z|x)||p_\theta(z)) + \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x|z^{(l)})
\end{aligned}
$$

Thus $\max ELBO = \min -ELBO = \min KL(q_\phi(z|x)||p_\theta(z)) + ||x-\hat{x}||^2$

How to calculate $KL(q_\phi(z|x)||p_\theta(z))$?

Appendix One dimension: $z \in \mathbb{R}$

$$
\begin{aligned}
q_\phi(z|x) &= \mathcal{N}(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]} \\
\mathbb{E}_{z\sim q_\phi(\cdot|x)}(z) &= \mu, \mathbb{E}_{z\sim q_\phi(\cdot|x)}(z^2) = \mu^2 + \sigma^2 \\
p_\theta(z) &= \mathcal{N}(0,1) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right] \\
KL(q_\phi(z|x)||p_\theta(z)) &= \mathbb{E}_{z\sim q_\phi(\cdot|x)}\left[\log \frac{q_\phi(z|x)}{p_\theta(z)}\right] \\
&= \mathbb{E}_{z\sim q_\phi(\cdot|x)} \log \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]}}{\frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right]} \\
&= \mathbb{E}_{z\sim q_\phi(\cdot|x)}[-\log\sigma - \frac{(z-\mu)^2}{2\sigma^2} + \frac{z^2}{2}] \\
&= -\frac{1}{2}[\log\sigma^2 + 1 - \mu^2 - \sigma^2]
\end{aligned}
$$

Appendix Multi dimension: $\boldsymbol{z} \in \mathbb{R}^J$

$$
\begin{aligned}
\int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} &= \int \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) \log \mathcal{N}(\boldsymbol{z}; 0, \boldsymbol{I}) d\boldsymbol{z} \\
&= -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} (\mu_j^2 + \sigma_j^2) \\
\int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} &= \int \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) \log \mathcal{N}(\boldsymbol{z}; \mu, \sigma^2) d\boldsymbol{z} \\
&= -\frac{J}{2} \log(2\pi) - \frac{1}{2} \sum_{j=1}^{J} (1 + \log \sigma_j^2)
\end{aligned}
$$

Appendix Multi dimension: $\boldsymbol{z} \in \mathbb{R}^J$

$$
\begin{aligned}
KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z})) &= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p_\theta(\boldsymbol{z})}\right] \\
&= \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\log q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] - \mathbb{E}_{\boldsymbol{z}\sim q_\phi(\cdot|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{z})\right] \\
&= \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log q_\phi(\boldsymbol{z}|\boldsymbol{x}) d\boldsymbol{z} - \int q_\phi(\boldsymbol{z}|\boldsymbol{x}) \log p_\theta(\boldsymbol{z}) d\boldsymbol{z} \\
&= -\frac{1}{2}\sum_{j=1}^{J}(1 + \log \sigma_j^2 - \mu_j^2 - \sigma_j^2)
\end{aligned}
$$