Auto-Encoding Variational Bayes/ Variational Graph Auto-Encoders

DSAIL @ KAIST

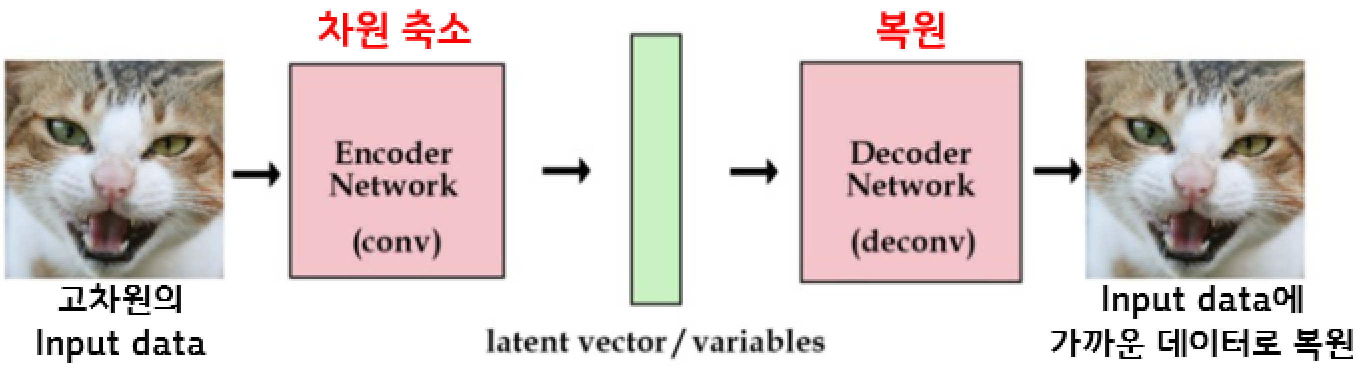
김원중

E-mail: wjkim@kaist.ac.kr

Contents

Auto Encoder

Auto-Encoder



Encoding 목적

Data가 가진 수십, 수백개의 변수로부터 정말 중요한 몇가지의 변수를 extraction하는 것

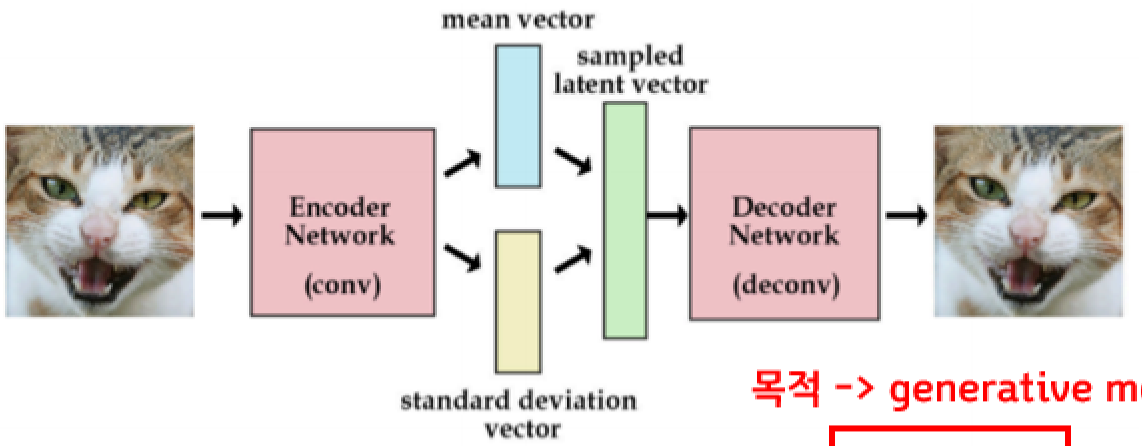
Decoding

다시 입력 data에 가까운 고차원 데이터로 복원

https://di-bigdata-study.tistory.com/4

Variational Auto Encoder

Variational Auto-Encoder



목적 -> generative model

Encoding

Decoder에 어떤 입력
latent variable z를 넣을지 학습

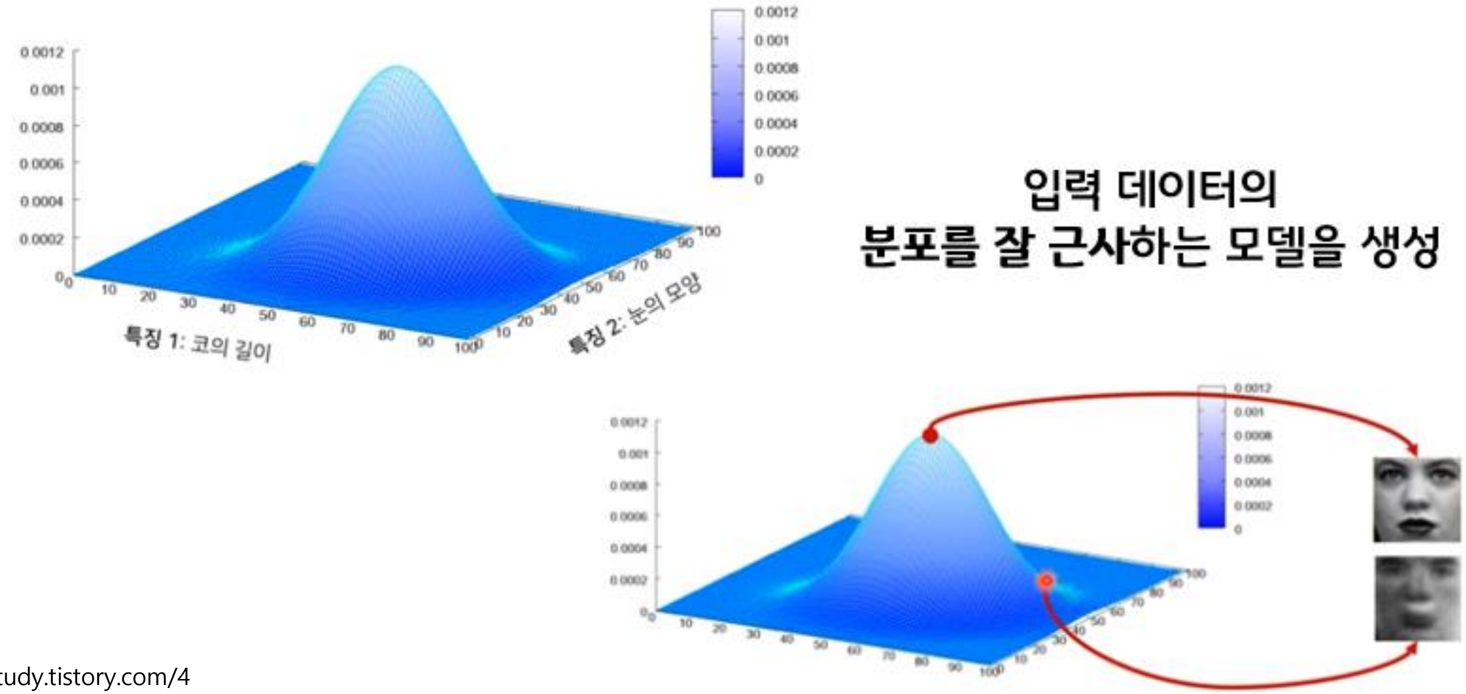
Decoding

입력데이터와 유사한
새로운 데이터 생성

- Purpose : Efficiently approximate inference and learning with directed probabilistic models whose continuous latent variables and/or parameters have intractable posterior distributions

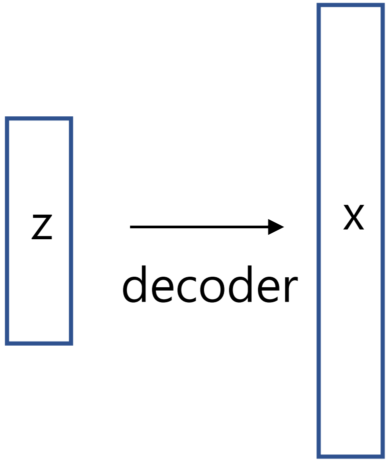
Generative model

- Generative model
  - Generative model can generate new data instance
  - Discriminative models discriminate between different kinds of data instances
  - Generative models capture the joint probability p(X, Y), or just p(X) if there are no label



https://di-bigdata-study.tistory.com/4

Problem scenario

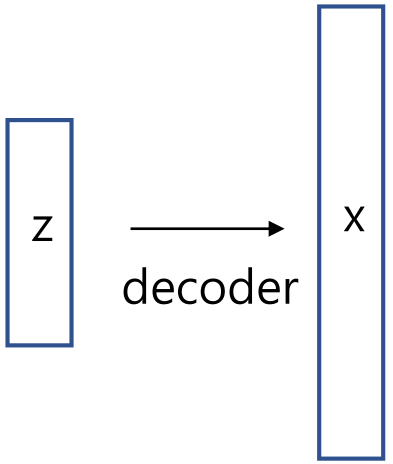


$p_{\boldsymbol{\theta}}(\mathbf{z})$ $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

Want to Maximize $p_{\boldsymbol{\theta}}(\mathbf{x})$

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

Problem scenario

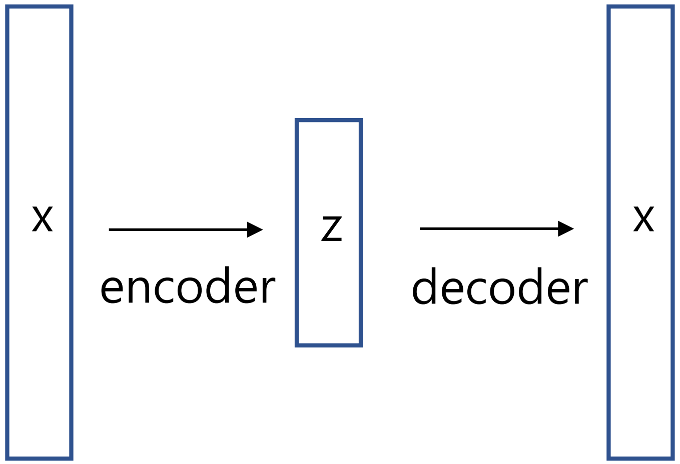


$p_{\boldsymbol{\theta}}(\mathbf{z})$ $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$.

Want to Maximize $p_{\boldsymbol{\theta}}(\mathbf{x})$

$$p_\theta(x) = \int p_\theta(z)\boxed{p_\theta(x|z)dz}$$

Problem scenario



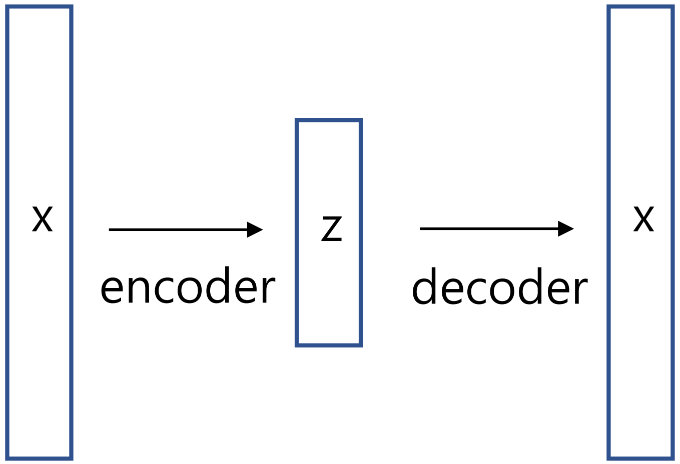
$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$

$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$

Want to Maximize $p_{\boldsymbol{\theta}}(\mathbf{x})$

$$p_\theta(x) = \int \boxed{p_\theta(z)} p_\theta(x|z)dz$$

Problem scenario



$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})$ intractable

$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$

$q_{\phi}(\mathbf{z}|\mathbf{x})$ tractable

Want to Maximize $p_{\boldsymbol{\theta}}(\mathbf{x})$

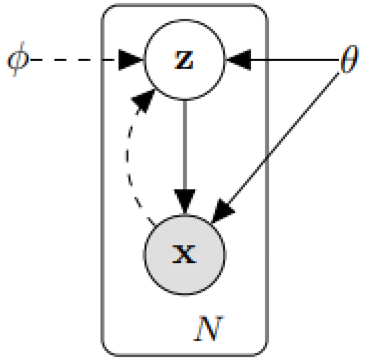
$$p_\theta(x) = \int \boxed{p_\theta(z)} p_\theta(x|z)dz$$

Problem scenario

- Dataset

$\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ Consisting of N i.i.d. samples of variable x

1) a value $\mathbf{z}^{(i)}$ is generated from prior distribution $p_{\theta^*}(\mathbf{z})$
2) a value $\mathbf{x}^{(i)}$ is generated from some conditional distribution $p_{\theta^*}(\mathbf{x}|\mathbf{z})$



→ Generative model $p_{\theta}(\mathbf{z})p_{\theta}(\mathbf{x}|\mathbf{z})$

⇢ Variational approximation $q_{\phi}(\mathbf{z}|\mathbf{x})$ to the intractable posterior $p_{\theta}(\mathbf{z}|\mathbf{x})$

Problem scenario

- Even works efficiently in the case of:
  1. Intractability
  2. A large dataset

- Propose solution to:
  1. 파라미터 θ에 대한 효율적인 근사 ML, MAP 추정
  2. 모수 θ 하에 관측된 변수 x의 값이 주어졌을 때 잠재변수 z에 대한 효율적인 근사 사후추론
  3. x에 대한 효율적인 marginal inference

Variational bound

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \underbrace{\boxed{D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)}))}}_{\geq 0} + \underbrace{\boxed{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})}}_{\text{lower bound}}$$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\right]$$

$$\log(p_\theta(x)) = \int_z q_\phi(z|x) \log(p_\theta(x)) \tag{1}$$

$$= \int_z q_\phi(z \mid x) \log \frac{p_\theta(z,\ x)}{p_\theta(z \mid x)} \tag{2}$$

$$= \int_z q_\phi(z \mid x) \log\left(\frac{p_\theta(z,\ x)}{q_\phi(z \mid x)} \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}\right) \tag{3}$$

$$= \boxed{\int_z q_\phi(z \mid x) \log \frac{p_\theta(z,\ x)}{q_\phi(z \mid x)}} + \boxed{\int_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}} \tag{4}$$

$$= \boxed{L(\theta,\ \phi;\ x)} + \boxed{D_{KL}(q_\phi(z \mid x) \parallel p_\theta(z \mid x))} \geq L(\theta,\ \phi;\ x) \tag{5}$$

$$KL(Q_\phi(Z|X)||P(Z|X)) = \sum_{z \in Z} q_\phi(z|x) \log \frac{q_\phi(z|x)}{p(z|x)}$$

Kullback Leibler divergence

https://ys-cs17.tistory.com/47

Variational bound

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)}) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) || p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$$

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \boxed{\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[-\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}) \right]}$$

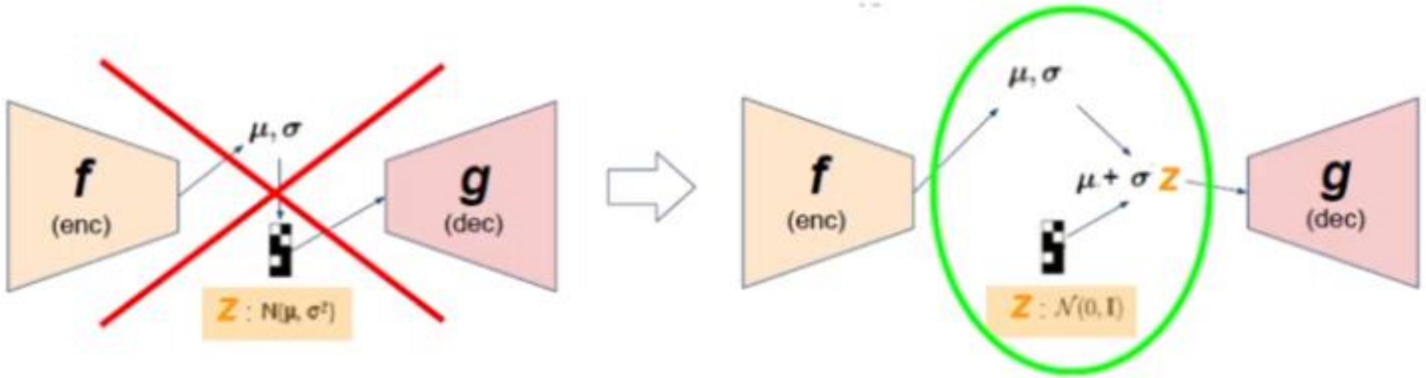
$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) || p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}) \right]$$

$$\begin{aligned}
\log(p_\theta(x)) &= \int_z q_\phi(z|x) \log(p_\theta(x)) && (1)\\
&= \int_z q_\phi(z \mid x) \log \frac{p_\theta(z,\ x)}{p_\theta(z \mid x)} && (2)\\
&= \int_z q_\phi(z \mid x) \log\left(\frac{p_\theta(z,\ x)}{q_\phi(z \mid x)} \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}\right) && (3)\\
&= \boxed{\int_z q_\phi(z \mid x) \log \frac{p_\theta(z,\ x)}{q_\phi(z \mid x)}} + \boxed{\int_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p_\theta(z \mid x)}} && (4)\\
&= \boxed{L(\theta,\ \phi;\ x)} + \boxed{D_{KL}(q_\phi(z \mid x) \parallel p_\theta(z \mid x))} \geq L(\theta,\ \phi;\ x) && (5)
\end{aligned}$$

$$\begin{aligned}
L(\theta,\ \phi;\ x) &= \int_z q_\phi(z \mid x) \log \frac{p_\theta(z,\ x)}{q_\phi(z \mid x)}\\
&= \int_z q_\phi(z \mid x) \log \frac{p_\theta(z) p_\theta(x \mid z)}{q_\phi(z \mid x)}\\
&= -\int_z q_\phi(z \mid x) \log \frac{q_\phi(z \mid x)}{p_\theta(z)} + \int_z q_\phi(z \mid x) \log p_\theta(x \mid z)\\
&= -D_{KL}(q_\phi(z \mid x) \| p_\theta(z)) + \mathbb{E}_{z|x}[\log p_\theta(x \mid z)]
\end{aligned}$$

Reparameterization trick

Reparameterization Trick



Sampling process $z^{i,l} \sim N(\mu_i, \sigma_i^2 I)$ ➡ $z^{i,l} = \mu_i + \sigma_i \odot \epsilon$

$\epsilon \sim N(0, I)$

Same distribution!
But it makes backpropagation possible!

SGVB(Stochastic Gradient Variational Bayes) estimator

$$\widetilde{\mathbf{z}} \sim q_{\phi}(\mathbf{z}|\mathbf{x}) \longrightarrow \widetilde{\mathbf{z}} = g_{\phi}(\boldsymbol{\epsilon}, \mathbf{x}) \quad \text{with} \quad \boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$$

Reparameterization trick

$$\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}\left[-\log q_{\phi}(\mathbf{z}|\mathbf{x}) + \log p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})\right] \longrightarrow$$

$$\widetilde{\mathcal{L}}^{A}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = \frac{1}{L}\sum_{l=1}^{L} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}, \mathbf{z}^{(i,l)}) - \log q_{\phi}(\mathbf{z}^{(i,l)}|\mathbf{x}^{(i)})$$

$$\text{where} \quad \mathbf{z}^{(i,l)} = g_{\phi}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)}) \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$$

Stochastic Gradient Variational Bayes (SGVB) estimator

$$\mathcal{L}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right] \longrightarrow$$

$$\widetilde{\mathcal{L}}^{B}(\boldsymbol{\theta}, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \frac{1}{L}\sum_{l=1}^{L}(\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)}))$$

$$\text{where} \quad \mathbf{z}^{(i,l)} = g_{\phi}(\boldsymbol{\epsilon}^{(i,l)}, \mathbf{x}^{(i)}) \quad \text{and} \quad \boldsymbol{\epsilon}^{(l)} \sim p(\boldsymbol{\epsilon})$$

another version of SGVB estimator

AEVB algorithm

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}) \simeq \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M) = \frac{N}{M} \sum_{i=1}^{M} \widetilde{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)})$$

Estimator of the lower bound of the full dataset

Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\boldsymbol{\epsilon} \leftarrow$ Random samples from noise distribution $p(\boldsymbol{\epsilon})$
  $\mathbf{g} \leftarrow \nabla_{\boldsymbol{\theta}, \boldsymbol{\phi}} \widetilde{\mathcal{L}}^M(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{X}^M, \boldsymbol{\epsilon})$ (Gradients of minibatch estimator (8))
  $\boldsymbol{\theta}, \boldsymbol{\phi} \leftarrow$ Update parameters using gradients $\mathbf{g}$ (e.g. SGD or Adagrad [DHS10])
until convergence of parameters $(\boldsymbol{\theta}, \boldsymbol{\phi})$
return $\boldsymbol{\theta}, \boldsymbol{\phi}$

Variational Auto-Encoder

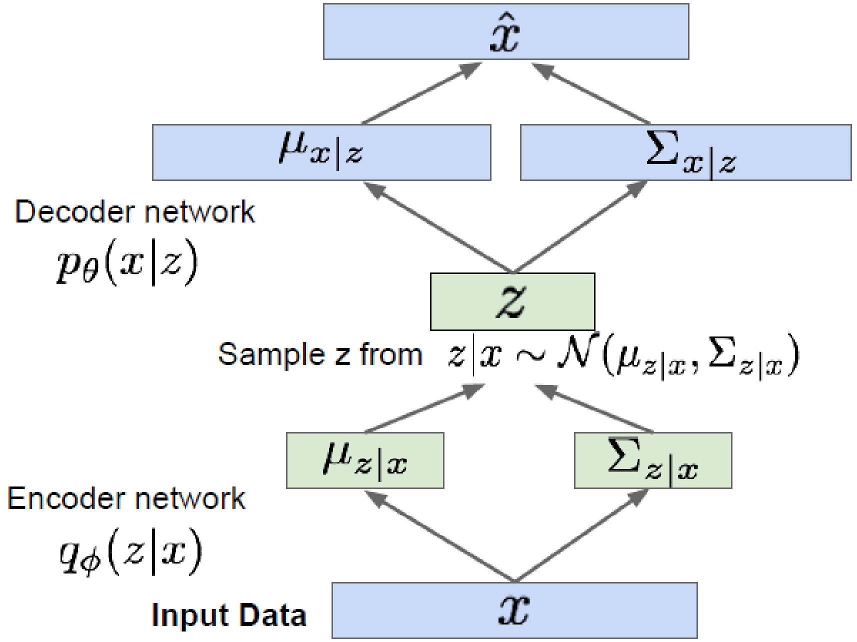
$$p_{\boldsymbol{\theta}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \mathbf{0}, \mathbf{I})$$

$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ multivariate gaussian (real-valued data) or Bernoulli (binary data)

$$\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) = \log \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}^{(i)}, \boldsymbol{\sigma}^{2(i)}\mathbf{I})$$

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) || p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right]$$

$$-D_{KL}((q_{\boldsymbol{\phi}}(\mathbf{z}) || p_{\boldsymbol{\theta}}(\mathbf{z})) = \int q_{\boldsymbol{\theta}}(\mathbf{z}) \left(\log p_{\boldsymbol{\theta}}(\mathbf{z}) - \log q_{\boldsymbol{\theta}}(\mathbf{z})\right) d\mathbf{z}$$
$$= \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log((\sigma_j)^2) - (\mu_j)^2 - (\sigma_j)^2\right)$$



$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}; \mathbf{x}^{(i)}) \simeq \frac{1}{2} \sum_{j=1}^{J} \left(1 + \log((\sigma_j^{(i)})^2) - (\mu_j^{(i)})^2 - (\sigma_j^{(i)})^2\right) + \frac{1}{L} \sum_{l=1}^{L} \log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z}^{(i,l)})$$

where $\mathbf{z}^{(i,l)} = \boldsymbol{\mu}^{(i)} + \boldsymbol{\sigma}^{(i)} \odot \boldsymbol{\epsilon}^{(l)}$ and $\boldsymbol{\epsilon}^{(l)} \sim \mathcal{N}(0, \mathbf{I})$ J : dimension of z

Lecture slide in CS231n, Stanford University

Variational Graph Auto-Encoder

Figure 1: Latent space of unsupervised VGAE model trained on Cora citation network dataset [1]. Grey lines denote citation links. Colors denote document class (not provided during training). Best viewed on screen.

- Definition

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N = |\mathcal{V}|$ nodes.

A : adjacency matrix
D : degree matrix
Z : N x F matrix (latent variables)
X : N x D matrix (node features)

Variational Graph Auto-Encoder

- Inference model

$$q(\mathbf{Z}\,|\,\mathbf{X},\mathbf{A}) = \prod_{i=1}^{N} q(\mathbf{z}_i\,|\,\mathbf{X},\mathbf{A}), \quad \text{with} \quad q(\mathbf{z}_i\,|\,\mathbf{X},\mathbf{A}) = \mathcal{N}(\mathbf{z}_i\,|\,\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2))$$

$\boldsymbol{\mu} = \mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X},\mathbf{A})$ is the matrix of mean vectors $\boldsymbol{\mu}_i$

$\log\boldsymbol{\sigma} = \mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{A})$

$$\mathrm{GCN}(\mathbf{X},\mathbf{A}) = \tilde{\mathbf{A}}\,\mathrm{ReLU}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}_0)\mathbf{W}_1$$

$\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix.

- Generative model

$$p(\mathbf{A}\,|\,\mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(A_{ij}\,|\,\mathbf{z}_i,\mathbf{z}_j), \quad \text{with} \quad p(A_{ij}=1\,|\,\mathbf{z}_i,\mathbf{z}_j) = \sigma(\mathbf{z}_i^\top\mathbf{z}_j)$$

where A_{ij} are the elements of $\mathbf{A}$ and $\sigma(\cdot)$ is the logistic sigmoid function

- Learning

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{A})}\big[\log p(\mathbf{A}\,|\,\mathbf{Z})\big] - \mathrm{KL}\big[q(\mathbf{Z}\,|\,\mathbf{X},\mathbf{A})\,||\,p(\mathbf{Z})\big]$$

$$p(\mathbf{Z}) = \prod_i p(\mathbf{z_i}) = \prod_i \mathcal{N}(\mathbf{z}_i\,|\,0,\mathbf{I})$$

Variational Graph Auto-Encoder

- Inference model

$$q(\mathbf{Z}\,|\,\mathbf{X},\mathbf{A}) = \prod_{i=1}^{N} q(\mathbf{z}_i\,|\,\mathbf{X},\mathbf{A}), \quad \text{with} \quad q(\mathbf{z}_i\,|\,\mathbf{X},\mathbf{A}) = \mathcal{N}(\mathbf{z}_i\,|\,\boldsymbol{\mu}_i, \mathrm{diag}(\boldsymbol{\sigma}_i^2))$$

$\boldsymbol{\mu} = \mathrm{GCN}_{\boldsymbol{\mu}}(\mathbf{X},\mathbf{A})$ is the matrix of mean vectors $\boldsymbol{\mu}_i$

$\log\boldsymbol{\sigma} = \mathrm{GCN}_{\boldsymbol{\sigma}}(\mathbf{X},\mathbf{A})$

$$\mathrm{GCN}(\mathbf{X},\mathbf{A}) = \tilde{\mathbf{A}}\,\mathrm{ReLU}(\tilde{\mathbf{A}}\mathbf{X}\mathbf{W}_0)\mathbf{W}_1$$

$\tilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$ is the symmetrically normalized adjacency matrix.

- Generative model

$$p(\mathbf{A}\,|\,\mathbf{Z}) = \prod_{i=1}^{N}\prod_{j=1}^{N} p(A_{ij}\,|\,\mathbf{z}_i,\mathbf{z}_j), \quad \text{with} \quad p(A_{ij}=1\,|\,\mathbf{z}_i,\mathbf{z}_j) = \sigma(\mathbf{z}_i^\top\mathbf{z}_j)$$

where A_{ij} are the elements of $\mathbf{A}$ and $\sigma(\cdot)$ is the logistic sigmoid function

- Learning

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X},\mathbf{A})}\big[\log p(\mathbf{A}\,|\,\mathbf{Z})\big] - \mathrm{KL}\big[q(\mathbf{Z}\,|\,\mathbf{X},\mathbf{A})\,||\,p(\mathbf{Z})\big]$$

$$p(\mathbf{Z}) = \prod_i p(\mathbf{z_i}) = \prod_i \mathcal{N}(\mathbf{z}_i\,|\,0,\mathbf{I})$$

$$p_{\boldsymbol{\theta}}(\mathbf{z}) = \mathcal{N}(\mathbf{z};\mathbf{0},\mathbf{I})$$

$p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ multivariate gaussian

$$\log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)}) = \log\mathcal{N}(\mathbf{z};\boldsymbol{\mu}^{(i)},\boldsymbol{\sigma}^{2(i)}\mathbf{I})$$

$$\mathcal{L}(\boldsymbol{\theta},\boldsymbol{\phi};\mathbf{x}^{(i)}) = -D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\boldsymbol{\theta}}(\mathbf{z})) + \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}^{(i)})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}|\mathbf{z})\right]$$

Variational Auto Encoder

Future work

- Better-suited prior distributions
- More flexible generative models
- Application of a SGD algorithm for improved scalability

End of Documents