Georgia Tech College of Computing

CS4641 Spring 2025

Latent Variable Model:

Variational Auto-Encoder

Bo Dai
School of CSE, Georgia Tech
bodai@cc.gatech.edu

Supervised Learning vs. Unsupervised Learning

Training Data

Supervised $\{x^i, y^i\}_{i=1}^m$

Unsupervised $\{x_i\}_{i=1}^m$

Learning Algorithm

Target Function

$f : X \to Y$

Density Estimation $p(x)$

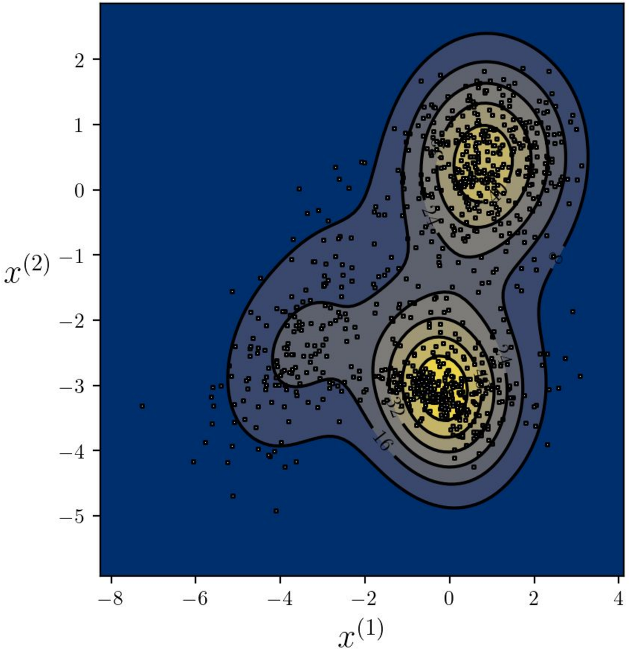
Clustering $f : X \to Y$

$Y \in \{0, 1, \ldots, k\}$

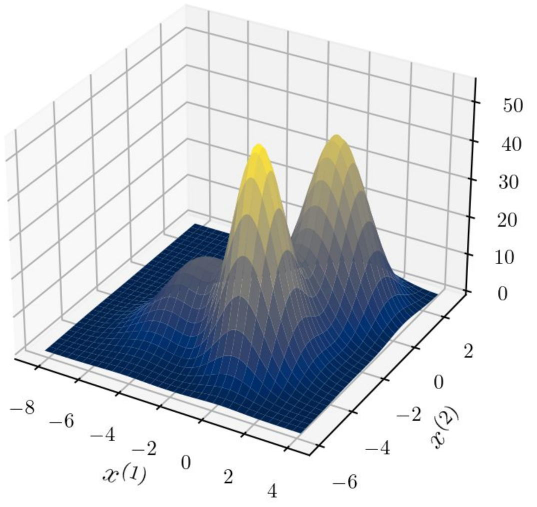
Dimension Reduction $f : X \to Y$

$X \in \mathbb{R}^p, \quad Y \in \mathbb{R}^d$

Density Estimation



$\{x_i\}_{i=1}^{m}$

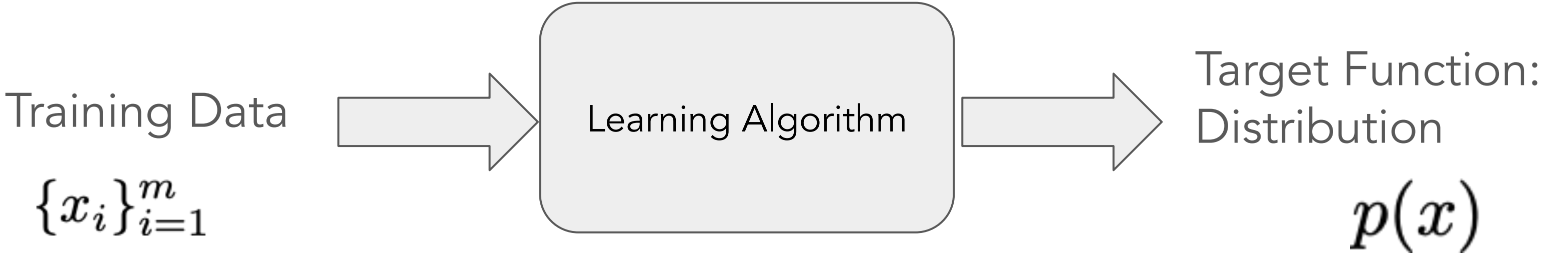


$p(x)$

Generative Models

$x \sim p(x)$

Density Estimation: Gaussian Mixture Model



Density Estimation Pipeline

1. Build probabilistic models
   Gaussian Mixture Model
2. Derive loss function (by MLE or MAP….)
   MLE
3. Select optimizer
   EM

Gaussian Mixture Model

Class mixture prior: $P(y)$ $\quad \pi = (\pi_1, \pi_2, \ldots, \pi_k), \quad \sum_{i=1}^{k} \pi_i = 1, \pi_i \geq 0$

Class conditional distribution: $p(x|y) = \mathcal{N}(x|\mu_y, \Sigma_y)$

Marginal distribution: $P(x) = \sum_y P(x|y)P(y) = \sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$

Expectation-Maximization

For t = 1……

- E-Step: Guess sample labels based on current model

$$\tau_j^l = \frac{\pi_l \mathcal{N}(x_j|\mu_l, \Sigma_l)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_j|\mu_l, \Sigma_l)}$$

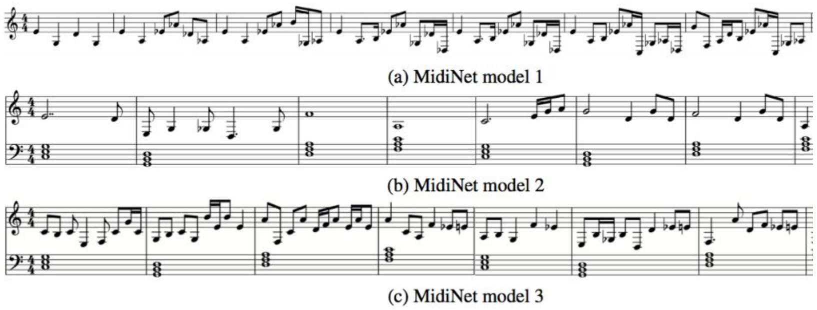
- M-Step: Update the parameters with current labels (Gaussian-Naive Bayes)

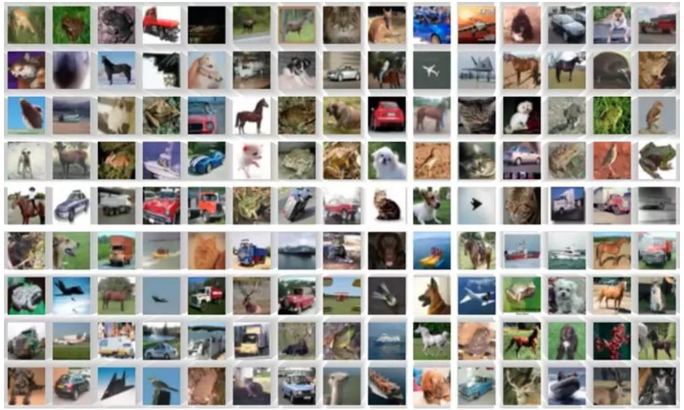
$$\mu_k = \frac{\sum_{i=1}^{m} \tau_k^i x^i}{\sum_{i=1}^{m} \tau_k^i}, \quad \pi_k = \frac{\sum_{i=1}^{m} \tau_k^i}{m}, \quad \Sigma_k = \frac{\sum_{i=1}^{m} \tau_k^i (x^i - \mu_k)(x^i - \mu_k)^\top}{\sum_{i=1}^{m} \tau_k^i}$$

This procedure is actually optimizing an upper bound of MLE, therefore, it converges

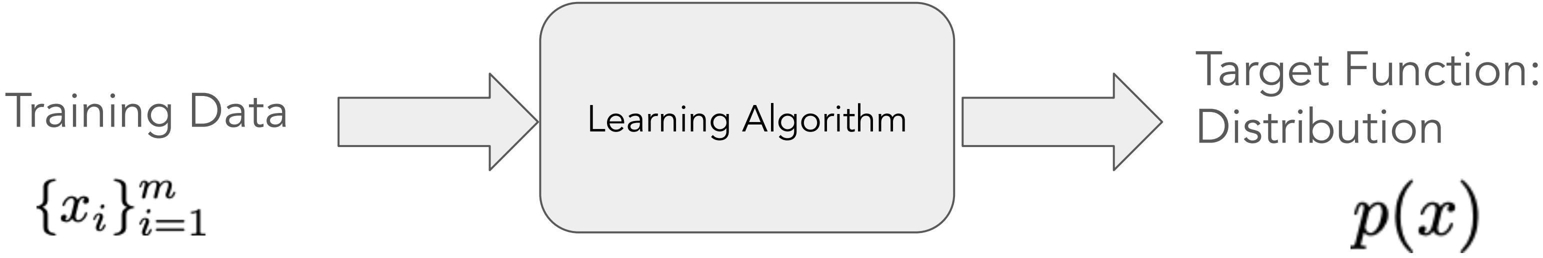
Density Estimation: Generative Models

$$x \sim p(x)$$





Generative Model: Latent Variable Models



Density Estimation Pipeline

1. Build probabilistic models
   Deep Latent Variable Model
2. Derive loss function (by MLE or MAP....)
3. Select optimizer

Latent Variable Models

GMMs $$\sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

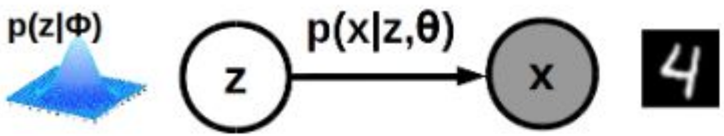
Latent Variable Models

Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

GMMs $$\sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

Not flexible enough

Latent Variable Models



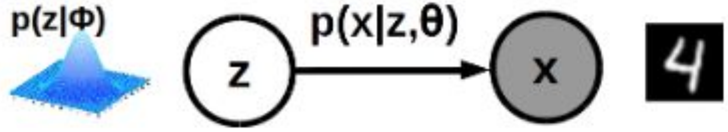
$$p(x) = \int p(x|z)p(z)dz$$

Figure 5: 1024 × 1024 images generated using the CELEBA-HQ dataset. See Appendix F for a larger set of results, and the accompanying video for latent space interpolations.

GMMs $\sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$

Not flexible enough

Latent Variable Models



$$p(x) = \int p(x|z)p(z)dz$$

Infinite-many components

Make $p(x|z)$ more flexible

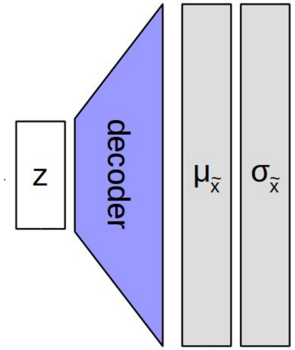
Deep Gaussian Distribution

Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}_z)}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1} (\mathbf{x} - \boldsymbol{\mu}_z)\Big)$$

Deep Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z)) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}(z))}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}(z))^T \boldsymbol{\Sigma}(z)^{-1} (\mathbf{x} - \boldsymbol{\mu}(z))\Big)$$



$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}\Big(\mathbf{x} \,\Big|\, \mu_{\tilde{x}}(z),\ \operatorname{diag}\big(\sigma^2_{\tilde{x}}(z)\big)\Big).$$

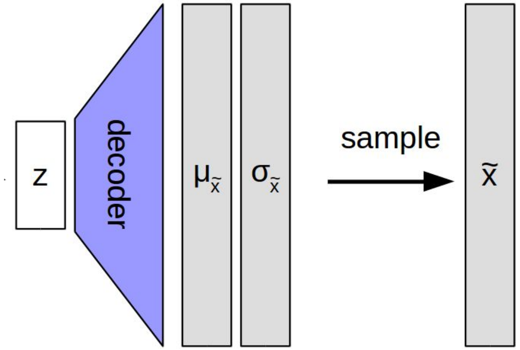
Deep Gaussian Distribution

Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}_z)}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1}(\mathbf{x} - \boldsymbol{\mu}_z)\right)$$

Deep Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z)) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}(z))}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}(z))^T \boldsymbol{\Sigma}(z)^{-1}(\mathbf{x} - \boldsymbol{\mu}(z))\right)$$



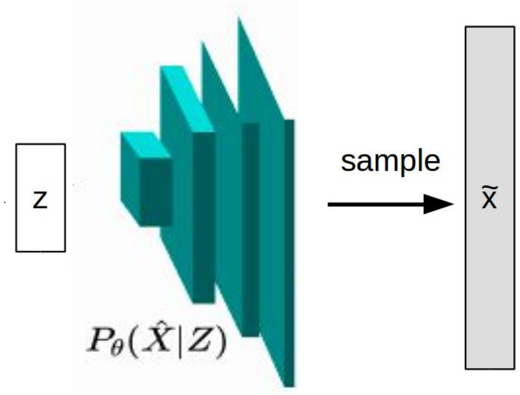
Deep Gaussian Distribution

Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}_z)}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1} (\mathbf{x} - \boldsymbol{\mu}_z)\Big)$$

Deep Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z)) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}(z))}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}(z))^T \boldsymbol{\Sigma}(z)^{-1} (\mathbf{x} - \boldsymbol{\mu}(z))\Big)$$



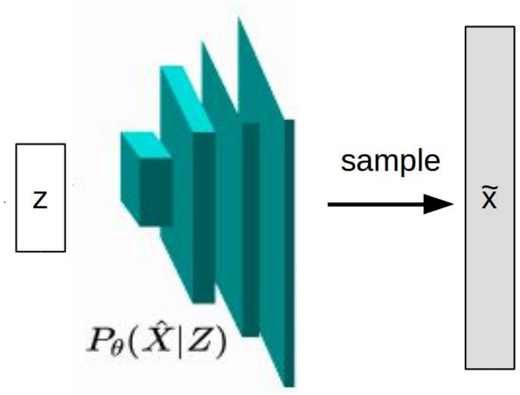
Deep Gaussian Distribution

Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}_z)}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1} (\mathbf{x} - \boldsymbol{\mu}_z)\Big)$$

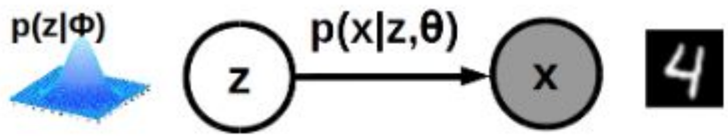
Deep Gaussian Distribution

$$p(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z)) = \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}(z))}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}(z))^T \boldsymbol{\Sigma}(z)^{-1} (\mathbf{x} - \boldsymbol{\mu}(z))\Big)$$



deep neural network $\mu_{W_\mu}(z), \sigma_{W_\sigma}(z)$

Deep Latent Variable Models: Deep Gaussian LVM



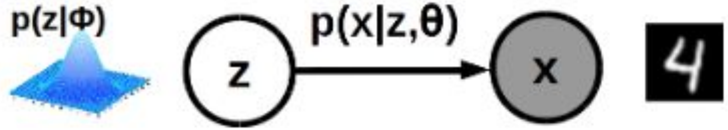
$$x \sim p(x) = \int p(x|z)p(z)dz$$

$$p(z) = \mathcal{N}(0, \sigma I)$$

$$\begin{aligned} p(\mathbf{x} \mid \mathbf{z}) &= \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z)) \\ &= \frac{1}{\sqrt{(2\pi)^d \det(\boldsymbol{\Sigma}(z))}} \exp\Big(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}(z))^T \boldsymbol{\Sigma}(z)^{-1}(\mathbf{x} - \boldsymbol{\mu}(z))\Big) \end{aligned}$$

Model Parameters $\mu_{W_\mu}(z), \sigma_{W_\sigma}(z) \;\; \sigma$

Sampling as Generation



$$x \sim p(x) = \int p(x|z)p(z)dz$$

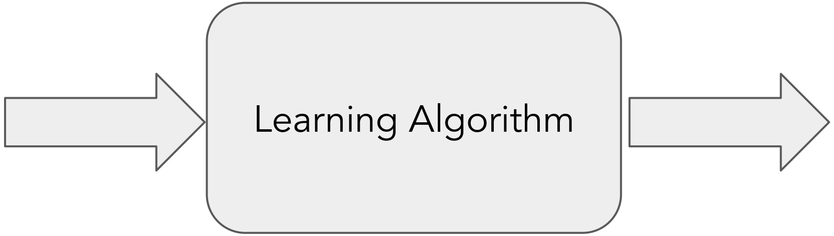
$$z \sim p(z) = \mathcal{N}(0, \sigma I)$$

$$x|z \sim \mathcal{N}(\mu(z), \sigma(z)I)$$

Generative Model: Latent Variable Models

Training Data

$\{x_i\}_{i=1}^m$



Target Function:
Distribution

$p(x)$

Density Estimation Pipeline

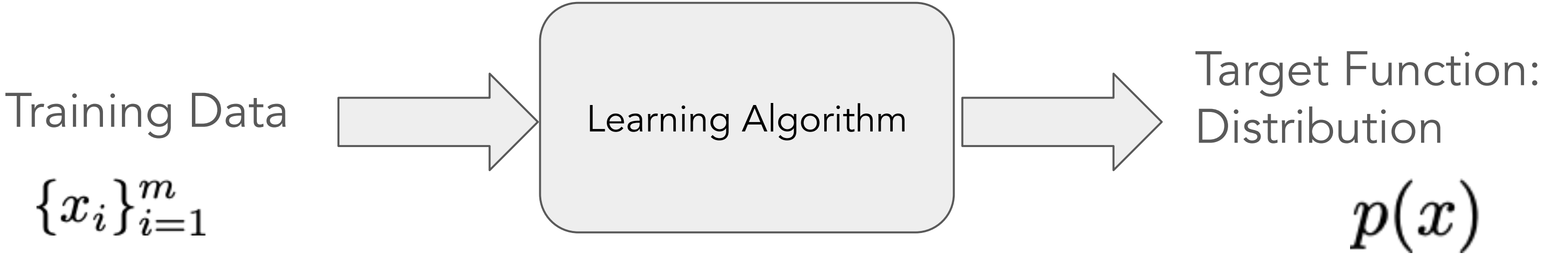
1. Build probabilistic models
2. Derive loss function (by MLE or MAP….)
   MLE?
3. Select optimizer

MLE of Deep LVM

$$p(x) = \int p(x|z)p(z)dz$$

$$\max_{\sigma, W_\mu, W_\sigma} \sum_{i=1}^{m} \log p(x^i) = \sum_{i=1}^{m} \log \boxed{\int p(z)p(x^i|z)dz}$$

Generative Model: Latent Variable Models



Density Estimation Pipeline

1. Build probabilistic models
2. Derive loss function (by MLE or MAP….)
   MLE Approximation: Evidence Lower BOund (ELBO) of MLE
3. Select optimizer

Recall GMMs

$$\sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

- E-Step:

$$\tau_j^l = \frac{\pi_l \mathcal{N}(x_j|\mu_l, \Sigma_l)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_j|\mu_l, \Sigma_l)}$$

- M-Step:

$$\max_{\pi_z, \mu_z, \Sigma_z} \sum_{i=1}^{m} \sum_{j=1}^{k} \tau_j^i \log \pi_j - \sum_{i=1}^{m} \log Z - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \tau_j^i (x^i - \mu_j)^\top \Sigma_j^{-1} (x^i - \mu_j)$$

Recall GMMs

$$\sum_{i=1}^{k} \pi_i \mathcal{N}(x|\mu_i, \Sigma_i)$$

- E-Step:

$$\tau_j^l = \frac{\pi_l \mathcal{N}(x_j|\mu_l, \Sigma_l)}{\sum_{l=1}^{k} \pi_l \mathcal{N}(x_j|\mu_l, \Sigma_l)} = q(y^i = j|x^i)$$

- M-Step:

$$\max_{\pi_z, \mu_z, \Sigma_z} \sum_{i=1}^{m} \sum_{j=1}^{k} \tau_j^i \log \pi_j - \sum_{i=1}^{m} \log Z - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{k} \tau_j^i (x^i - \mu_j)^\top \Sigma_j^{-1} (x^i - \mu_j)$$

$$\max_{\pi_z, \mu_z, \Sigma_z} \sum_{i=1}^{m} \sum_{j=1}^{k} q(y^i = j|x^i) \left(\log p(x^i, \tau^i)\right)$$

Intuitive Idea

$$p(x) = \int p(x|z)p(z)dz$$

- E-Step:

  Calculate $q(z^i|x^i)$

- M-Step:

$$\max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\Big(\log p(x^i|z^i)p(z^i) \Big) dz^i$$

Intuitive Idea

$$p(x) = \int p(x|z)p(z)dz$$

- E-Step:

Calculate $q(z^i|x^i) = \dfrac{p(z^i)p(x^i|z^i)}{\boxed{\int p(z^i)p(x^i|z^i)dz^i}}$

- M-Step:

$$\max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\Big(\log p(x^i|z^i)p(z^i)\Big)dz^i$$

Revisit K-means

- K-means Objective

$$\min_{\{\boldsymbol{\mu}\},\boxed{\{\mathbf{y}\}}} J(\{\boldsymbol{\mu}\},\{\mathbf{y}\}) = \min_{\{\boldsymbol{\mu}\},\{\mathbf{y}\}} \sum_{n=1}^{N}\sum_{k=1}^{K} y_k^{(n)} \|\boldsymbol{\mu}_k - \mathbf{x}^{(n)}\|^2$$

$$\text{s.t. } \sum_k y_k^{(n)} = 1, \forall n, \text{ where } y_k^{(n)} \in \{0,1\}, \forall k, n$$

Intuitive Idea

$$p(x) = \int p(x|z)p(z)dz$$

$$\boxed{\max_{q(z|x)}} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\Big(\log p(x^i|z^i)p(z^i) \Big) dz^i$$

Intuitive Idea

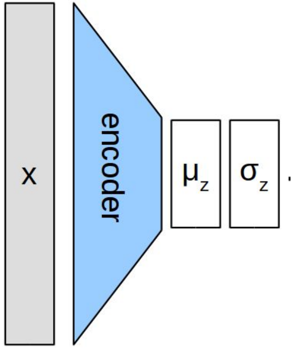
$$p(x) = \int p(x|z)p(z)dz$$

$$\boxed{\max_{q(z|x)}} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\Big(\log p(x^i|z^i)p(z^i)\Big)dz^i - \int q(z^i|x^i) \log q(z^i|x^i) dz_i$$

Intuitive Idea

$$p(x) = \int p(x|z)p(z)dz$$

$$\boxed{\max_{q(z|x)}} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\left(\log p(x^i|z^i)p(z^i) \right) dz^i - \int q(z^i|x^i) \log q(z^i|x^i) dz_i$$

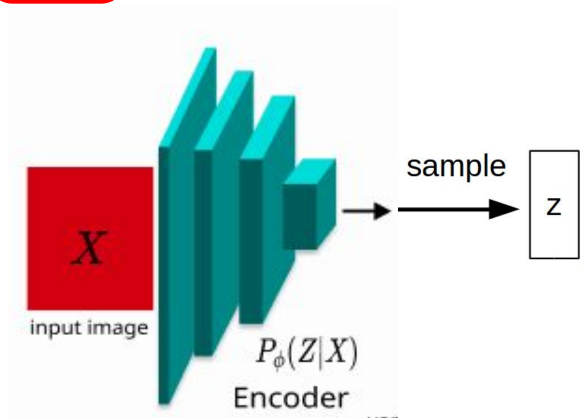


$$q(z|x) = \mathcal{N}\left(z \mid \mu_z(x), \mathrm{diag}(\sigma_z^2(x))\right)$$

Intuitive Idea

$$p(x) = \int p(x|z)p(z)dz$$

$$\boxed{\max_{q(z|x)}} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\Big(\log p(x^i|z^i)p(z^i)\Big)dz^i - \int q(z^i|x^i)\log q(z^i|x^i)dz_i$$



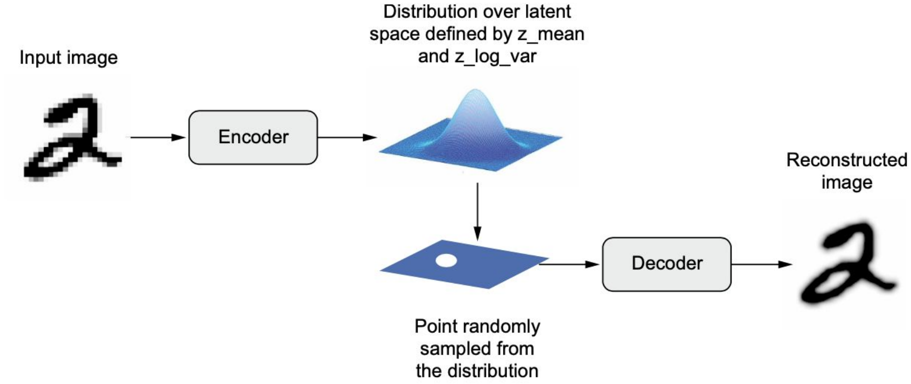
$$q(z|x) = \mathcal{N}\left(z \mid \mu_z(x), \mathrm{diag}(\sigma_z^2(x))\right)$$

deep neural network

$$\mu_{W_z}(x), \quad \sigma_{W_z}(x)$$

Evidence Lower Bound

$$\boxed{\max_{W_z^1, W_z^2}} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\left(\log p(x^i|z^i)p(z^i)\right)dz^i \underbrace{- \int q(z^i|x^i)\log q(z^i|x^i)dz_i}_{H(q(z|x))}$$



Evidence Lower Bound

$$\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\left(\log p(x^i|z^i)p(z^i) \right) dz^i + H(q(z|x))$$

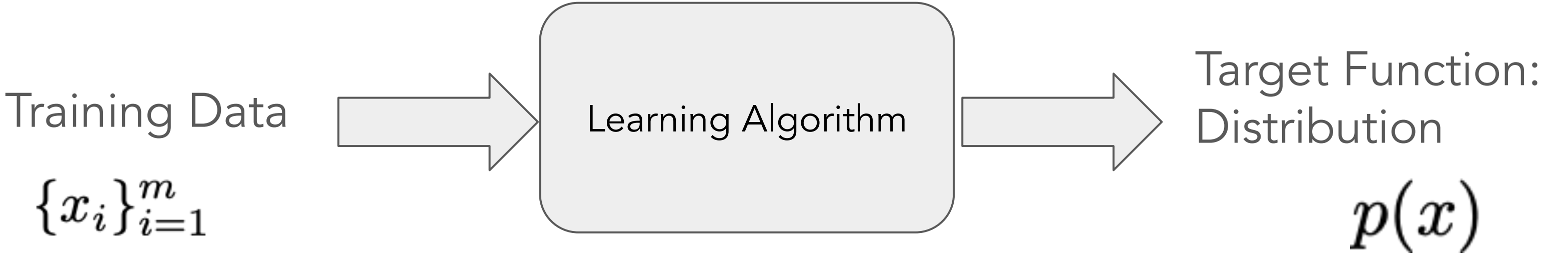
$$= \max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\left(\log \frac{p(x^i|z^i)p(z^i)}{q(z^i|x^i)} \right) dz^i$$

Evidence Lower Bound

$$
\begin{aligned}
&\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i) \Big(\log p(x^i|z^i) p(z^i) \Big) dz^i + H(q(z|x)) \\
&= \max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i) \left(\log \frac{p(x^i|z^i) p(z^i)}{q(z^i|x^i)} \right) dz^i \\
&\leq \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \log \int \left(\frac{p(x^i|z^i) p(z^i)}{q(z^i|x^i)} q(z^i|x^i) dz^i \right)
\end{aligned}
$$

$$\mathbb{E}[\log Y] \leq \log \mathbb{E}[Y]$$

Generative Model: Latent Variable Models



Density Estimation Pipeline

1. Build probabilistic models
2. Derive loss function (by MLE or MAP….)
3. Select optimizer

   Stochastic Gradient

Evidence Lower Bound

$$\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \int q(z^i|x^i)\left(\log p(x^i|z^i)p(z^i)\right) dz^i \underbrace{- \int q(z^i|x^i) \log q(z^i|x^i) dz_i}_{H(q(z|x))}$$

$$\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \mathbb{E}_{q(z^i|x^i)}\left[\log p(x^i|z^i)p(z^i) - \log q(z^i|x^i)\right]$$

Reparamerization Trick

$$\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \mathbb{E}_{q(z^i|x^i)} \left[\log p(x^i|z^i) p(z^i) - \log q(z^i|x^i) \right]$$

$$z \sim \; q(z|x) = \mathcal{N}\left(z \mid \mu_z(x), \mathrm{diag}(\sigma_z^2(x))\right)$$

Reparamerization Trick

$$\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \mathbb{E}_{q(z^i|x^i)} \left[\log p(x^i|z^i) p(z^i) - \log q(z^i|x^i) \right]$$

$$z \sim \; q(z|x) = \mathcal{N}\left(z \mid \mu_z(x), \operatorname{diag}(\sigma_z^2(x))\right)$$

$$z = \mu_z(x) + \sigma_z(x)\epsilon \qquad\qquad \epsilon \sim \mathcal{N}(0, I)$$

Reparamerization Trick

$$\max_{q(z|x)} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \mathbb{E}_{q(z^i|x^i)} \Big[\log p(x^i|z^i) p(z^i) - \log q(z^i|x^i) \Big]$$

$$z \sim \; q(z|x) = \mathcal{N}\left(z \mid \mu_z(x), \mathrm{diag}(\sigma_z^2(x))\right)$$

$$z = \mu_z(x) + \sigma_z(x)\epsilon \qquad\qquad \epsilon \sim \mathcal{N}(0, I)$$

$$\max_{W_z^1, W_z^2} \max_{\sigma, W_\sigma, W_\mu} \sum_{i=1}^{m} \mathbb{E}_{\epsilon \sim \mathcal{N}(0,I)} \Big[\log p(x^i|\mu(x^i) + \sigma(x^i)\epsilon) p(\mu(x^i) + \sigma(x^i)\epsilon)$$
$$- \log q(\mu(x^i) + \sigma(x^i)\epsilon | x^i) \Big]$$

Generative Model: Latent Variable Models

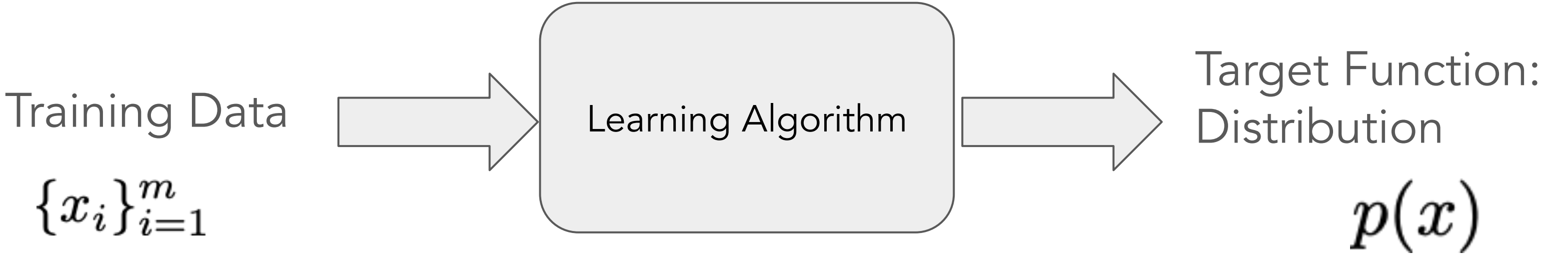
Training Data $\{x_i\}_{i=1}^m$ → Learning Algorithm → Target Function: Distribution $p(x)$

Density Estimation Pipeline

1. Build probabilistic models
   Deep Latent Variable Model
2. Derive loss function (by MLE or MAP….)
   ELBO
3. Select optimizer
   Stochastic Gradient

VAE Generation

Variants of VAE



Density Estimation Pipeline

1. Build probabilistic models

   Deep Latent Variable Model: Beyond Gaussian
2. Derive loss function (by MLE or MAP….)

   ELBO
3. Select optimizer

   Stochastic Gradient

Q&A