

LEC9 \rightarrow VAE

1. Generative Model

Motivation: Density Estimation

Given $\mathcal{D} = \{(x_i)\}_{i=1}^N$, $x_i \sim p^*$ (GT)

Goal: find (estimate) $\hat{p} \approx p^*$ (parametric model)

\Rightarrow Generative Model:

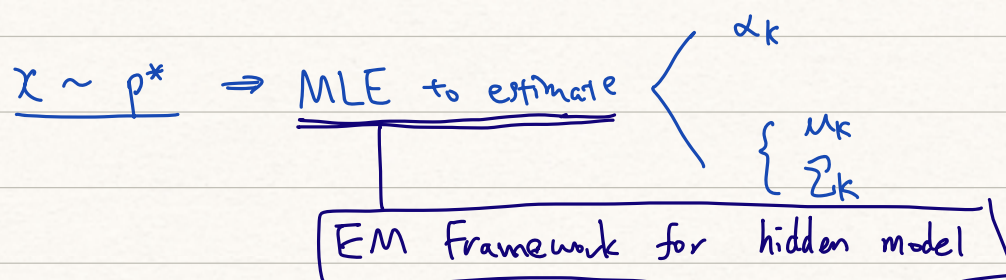
Given $\mathcal{D} = \{(x_i)\}_{i=1}^N$, $x_i \sim p^*$ (GT)

Goal sample $x^{(new)} \sim p^*$ approximately

[Toy Example] GMM

① $z \sim \text{Multi-bernoulli}(K) \in [K]$

② $x \sim \text{Gaussian}(\mu_z, \Sigma_z)$



2. Limitation of AE

① hidden space has no structure (not smooth w.r.t distribution of data points)

\downarrow
we should give some structure to hidden space

\Downarrow
example $z \sim \mathcal{N}(0, I_K)$

3. Latent Variable Model (GMM is one example)

$$P_{\theta}(x) = \int_{\mathcal{Z}} P_{\theta}(x|z) P_{\theta}(z) dz$$

$\mathcal{Z} \rightarrow$ Hidden (Latent) Variable

$\begin{cases} P_{\theta}(z) : \text{latent variable prior} \\ P_{\theta}(x|z) : \text{conditional distribution} \end{cases}$

Generating Process: model $\begin{cases} P_{\theta}(z) \\ P_{\theta}(x|z) \end{cases} \rightarrow \text{Learn from } \mathcal{D}$

\rightarrow Given $x^{(\text{new})}$

\rightarrow calculate $P_{\theta}(z|x^{(\text{new})})$ from Bayes

$\rightarrow z^{(\text{new})} \sim P_{\theta}(z|x^{(\text{new})})$ \rightarrow intractable might be

$\rightarrow x' \sim P_{\theta}(x|z^{(\text{new})})$

approximate with $q_{\phi}(\cdot)$
(Variational Inference)

4. VAE setting

$\begin{cases} \textcircled{1} z \sim \text{Gaussian}(\mathbf{0}, \mathbf{I}_K) \\ \textcircled{2} x|z \sim \text{Gaussian}(\mu_{\text{dec}}(z; \theta), \sigma_{\text{dec}}^2(z; \theta) \mathbf{I}_D) \end{cases}$

CE
(also can use Bernoulli model)

Square Loss

Note: (x, z) is not Gaussian any more

\Rightarrow intractable $z|x$!

Therefore, approximate $z|x$ via Variational Inference

$z|x \sim \text{Gaussian}(f_{\phi}(x) + g_{\phi}(x)^2 \mathbf{I}_K)$ Approximately

$\Rightarrow z|x = f_{\phi}(x) + g_{\phi}^2(x) \circ \varepsilon$ Here $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_K)$

\rightarrow reparameterize trick

Question : How to learn parameters ?

$$\begin{aligned} \textcircled{1} \quad \text{MLE} \quad \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} \log P(X|\theta) \Rightarrow \boxed{\text{log-likelihood } \mathcal{L}(\theta)} \\ &= \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^n \log \int_{\mathbf{z}} P(X_i | \mathbf{z}) P(\mathbf{z}) d\mathbf{z} \\ &\quad \text{(intractable)} \end{aligned}$$

Decomposition $\mathcal{L}(\theta)$:

$$\begin{aligned} \mathcal{L}(\theta) &= \log P(X|\theta) \\ &= \mathbb{E}_{\mathbf{z} \sim q} [\log P(X|\theta)] \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}} \left[\log \frac{P(X, \mathbf{z}|\theta)}{P(\mathbf{z}|X, \theta)} \right] \\ &= \mathbb{E}_{\mathbf{z} \sim q_{\phi}} \left[\log \frac{P(X, \mathbf{z}|\theta)}{q_{\phi}(\mathbf{z})} \right] - \mathbb{E}_{\mathbf{z} \sim q_{\phi}} \left[\log \frac{P(\mathbf{z}|X, \theta)}{q_{\phi}(\mathbf{z})} \right] \\ &= \underline{\text{ELBO}} + D_{\text{KL}}(q_{\phi}(\cdot) \| P(\cdot|X; \theta)) \end{aligned}$$

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

$$\textcircled{2} \quad \xrightarrow{\text{Maximize ELBO}} (\hat{\theta}, \hat{\phi}) = \underset{\theta, \phi}{\operatorname{argmax}} \text{ELBO}(\theta, \phi)$$

$$= \underset{\theta, \phi}{\operatorname{argmax}} \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\cdot)} \left[\log \frac{P(X, \mathbf{z}|\theta)}{q_{\phi}(\mathbf{z})} \right]$$

$$= \underset{\theta, \phi}{\operatorname{argmax}} \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\cdot)} [\log P(X|\mathbf{z}; \theta)]$$

$$+ \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\cdot)} \left[\frac{P(\mathbf{z}|\theta)}{q_{\phi}(\mathbf{z})} \right]$$

$$= \underset{\theta, \phi}{\operatorname{argmax}} \mathbb{E}_{\mathbf{z} \sim q_{\phi}(\cdot)} [\log P(X|\mathbf{z}; \theta)] - D_{\text{KL}}(q_{\phi}(\cdot) \| P(\cdot|\theta))$$

Reparameterize trick: $= \underset{\theta, \phi}{\operatorname{argmin}} - \mathbb{E}_{z \sim q_{\phi}(\cdot|x)} \left[\frac{1}{2} \|x - \delta(z)\|_2^2 \right]$

$\mathbb{E}_{z \sim q_{\phi}(\cdot|x)} [\cdot]$

$= \mathbb{E}_{u \sim \mathcal{N}(0, I_k)} [\cdot]$

$z = f_{\phi}(x) + g_{\phi}(x) \odot u$

reconstruction loss (decoder)

$+ D_{KL}(q_{\phi}(\cdot|x) \parallel p(\cdot|\theta))$

regularization term (encoder)

require sample

closed form

③ We will have $q_{\phi}(\cdot|x) \approx p(z|x)$

encoder part

Reason: given θ , $p(x|\theta)$ is fixed

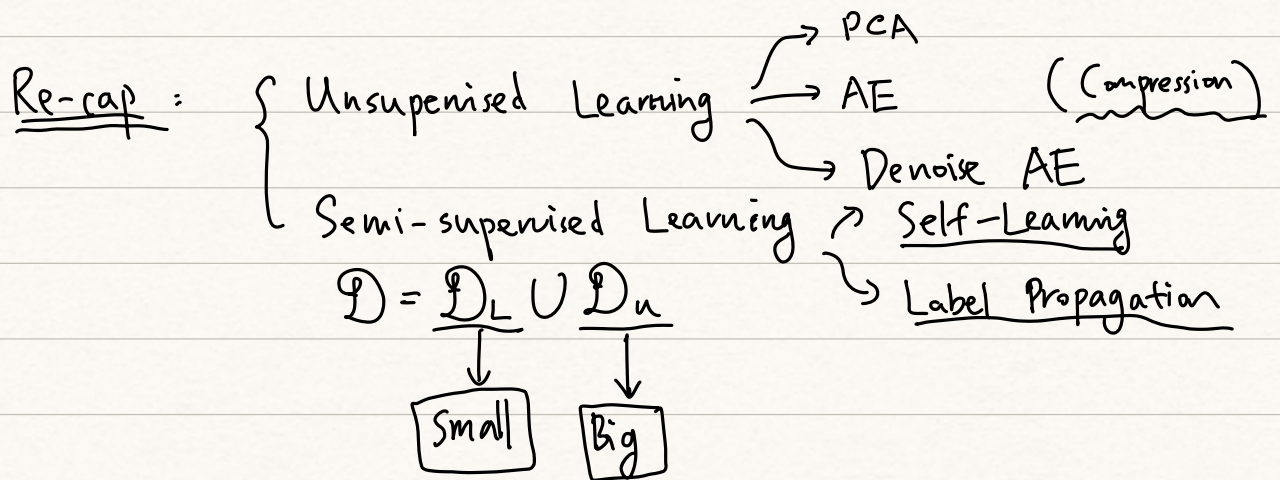
$$\log p(x|\theta) = \text{ELBO} + D_{KL}(q(\cdot|x) \parallel p(\cdot|x;\theta))$$

since $\hat{\phi} = \underset{\phi}{\operatorname{argmax}} \text{ELBO} \quad \text{given } \theta$

$$= \underset{\phi}{\operatorname{argmin}} D_{KL}(q_{\phi}(\cdot|x) \parallel p(\cdot|x;\theta))$$

\Rightarrow we can expect $q_{\hat{\phi}}(z|x) \approx p(z|x;\theta)$

Lec 9



Today :

1. Learning Across Tasks

↓
Assumption: Manifold Assumption

① Transfer Learning → { Data Similar
Task Very Different

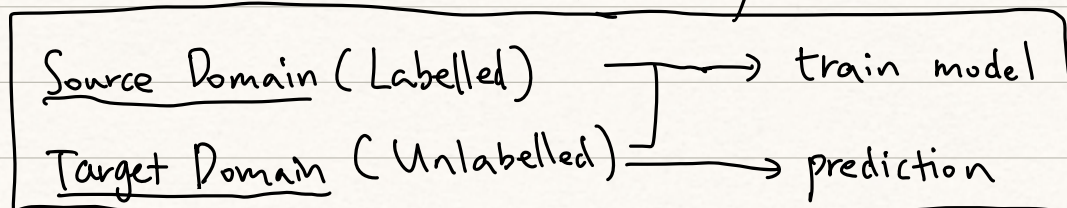
we hope we can learn some features that are useful for different tasks

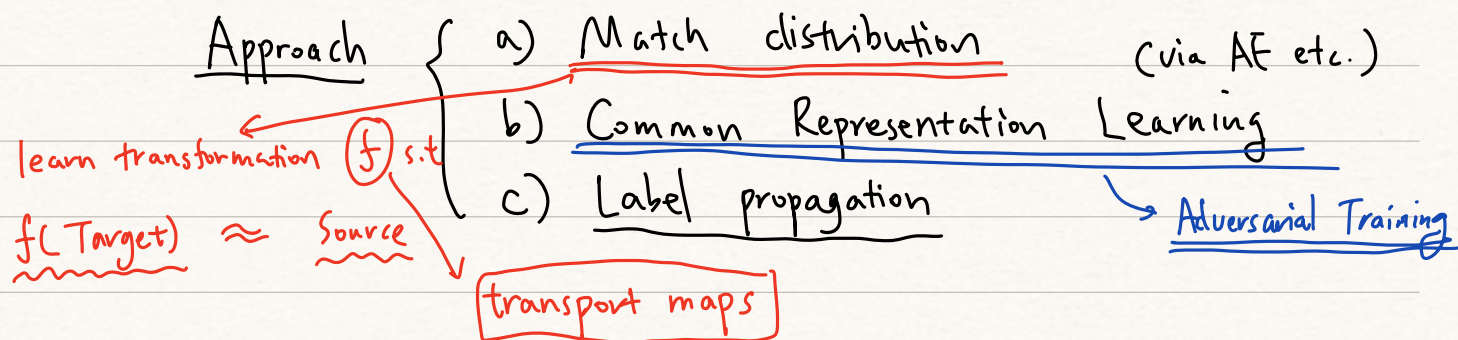
Approach { fine-tune & pre-train
warm-start

(w.r.t distribution)

② Domain Adaptation → { Data Source Different
Task Very Similar

Idea ⇒





2. VAE

① Probabilistic Modelling

- a) Density Estimation
- b) Generative Model

a) Density Estimation



Parametric density model $p_{\theta}(x)$

↓
MLE to estimate $\hat{\theta}$

$$\underline{x \sim p^* \approx p_{\hat{\theta}}}$$

aim is $\underline{p_{\hat{\theta}}}$

density function

↓
difficult to sample

b) Generative Model

$$\underline{x \sim p^*} \Rightarrow \text{approximately } \underline{\hat{x} \sim \hat{p}}$$

aim is sample \hat{x}

② GMM

$$\begin{cases} z \sim \text{Multi-Bernoulli}(\alpha) \\ x|z \sim \text{Gaussian}(\mu_z, \Sigma_z) \end{cases}$$

$$p_{\theta}(x) = \sum_i \alpha_i p_{\text{Gaussian}}(x; \mu_i, \Sigma_i)$$

⇒ Both Density Estimation & Generative Model

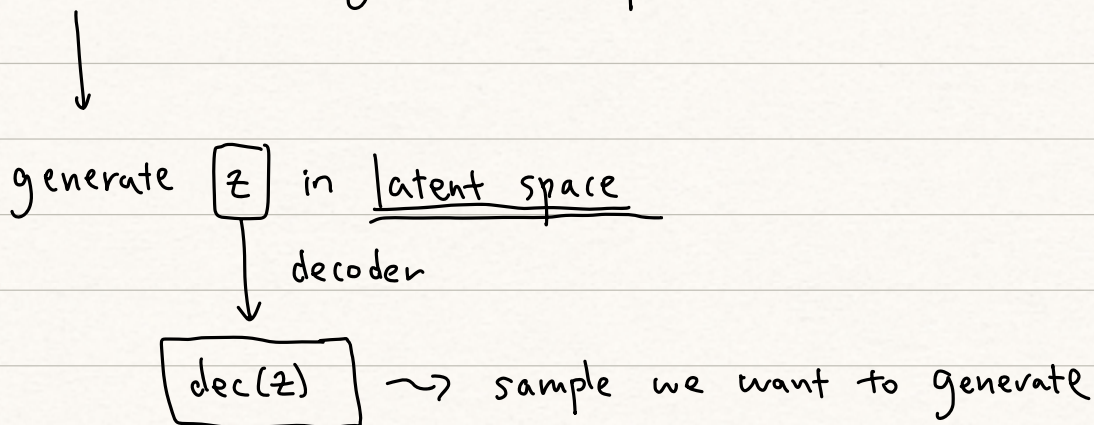
Learning is via MLE + EM Framework

$$\Theta = \{ \{ \hat{\alpha}_i \}_{i=1}^K, \{ \hat{\mu}_i, \hat{\Sigma}_i \}_{i=1}^K \}$$

Limitation:

- a) Not complex enough w.r.t model capacity
- b) Need to choose (# of clusters) → hyper-parameter
- c) No prior knowledge (translation invariance of CNN)

③ Utilize AE to generate samples?



④ To construct a "Generative Model"

we need { able to generate
model capacity

⇒ latent variable model

$$P_{\theta}(x) = \int_{\mathcal{Z}} P_{\theta}(x|z) P_{\theta}(z) dz$$

$\begin{cases} p_0(z) \rightarrow \text{prior dist} \\ p_0(x|z) \rightarrow \text{generative dist} \end{cases}$

Generating Scheme

1. learn $p_0(z)$ & $p_0(x|z)$

2. calculate $p_0(z|x)$ from Bayes Theorem (VI)

3. when we have x



sample from

$$z \sim p_0(z|x)$$

\approx

$$q_\phi(z|x)$$

intractable



sample from $\tilde{x} \sim p_0(x|z)$