Study on Latent Representation and Clustering

Kyeong Ryeol, Go
M.S. Candidate of OSI Lab

Explicit distribution

1. Disentanglement
    1. Beta-VAE
    2. Factor-VAE
    3. Beta-TC-VAE
    4. HSIC-constrained-VAE

Factors of variation이 latent space의 axis로 각각 align되도록 함.
i) 기존의 ELBO에 penalty term을 추가하거나,
ii) 기존의 ELBO를 decompose해서 특정 component의 weight를 세게 주는 방식이 있다.

2. Gaussian Mixture Model
    1. Variational Deep Embedding (VaDE)
    2. Gaussian Mixture VAE (GM-VAE)

3. Dirichlet Mixture Model
    1. Stick Breaking VAE (SB-VAE)
    2. Dirichlet VAE (Dir-VAE)

4. Flow-based
    1. Normalizing Flows (NF)
    2. Inverse Autoregressive Flows (IAF)

Disentanglement

1. Beta-VAE
   - $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot KL(q_\phi(z|x)||p(z))$
   - $\beta = 1 \;\rightarrow original\ ELBO$

2. Factor-VAE
   - $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot KL\left(q_\phi(z) \,\middle\|\, \prod_d q_\phi(z_d)\right) - KL(q_\phi(z|x)||p(z))$
     $(q_\phi(z) = E_{p(x)}[q_\phi(z|x)])$
   - $KL\left(q_\phi(z) \,\middle\|\, \prod_d q_\phi(z_d)\right) \approx E_{q_\phi(z)}\left[\log \frac{D(z)}{1-D(z)}\right]$ $where\ D\ is\ a\ discriminator : q_\phi(z)\ \ v.s.\ \prod_d q_\phi(z_d)$
     (Inner optimization loop exists)
   - $\beta = 0 \;\rightarrow original\ ELBO$

Disentanglement

3. Beta-TC-VAE

- $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot KL\left(q_\phi(z) \middle\| \prod_d q_\phi(z_d)\right) - KL\big(q_\phi(z,x) \| q_\phi(z)p(x)\big) - \sum_d KL\big(q_\phi(z_d) \| p(z_d)\big)$
- $KL\left(q_\phi(z) \middle\| \prod_d q_\phi(z_d)\right) = E_{q_\phi(z)}[\log q_\phi(z)] - E_{q_\phi(z)}\left[\log \prod_d q_\phi(z_d)\right]$
  - $E_{q_\phi(z)}[\log q_\phi(z)] \approx \frac{1}{M}\sum_{i=1}^{M} \log \sum_{j=1}^{M} \exp\big(\log \sum_d q_\phi(z_d^i | x^j)\big) - \log NM$
  - $E_{q_\phi(z)}\left[\log \prod_d q_\phi(z_d)\right] = \sum_d E_{q_\phi(z_d)}[\log q_\phi(z_d)] \approx \sum_d \frac{1}{M}\sum_{i=1}^{M} \log \sum_{j=1}^{M} \exp\big(\log q_\phi(z_d^i | x^j)\big) - \log NM$
- $\beta = 1 \ \rightarrow original\ ELBO$
  ($objective\ function\ can\ be\ identical\ to\ that\ of$ Factor-VAE)

4. HSIC-constrained-VAE

- $\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot HSIC\left(q_\phi(z)\right) - KL\big(q_\phi(z|x) \| p(z)\big)$
- $\beta = 0 \ \rightarrow original\ ELBO$

Measure between prob. dist. by kernel

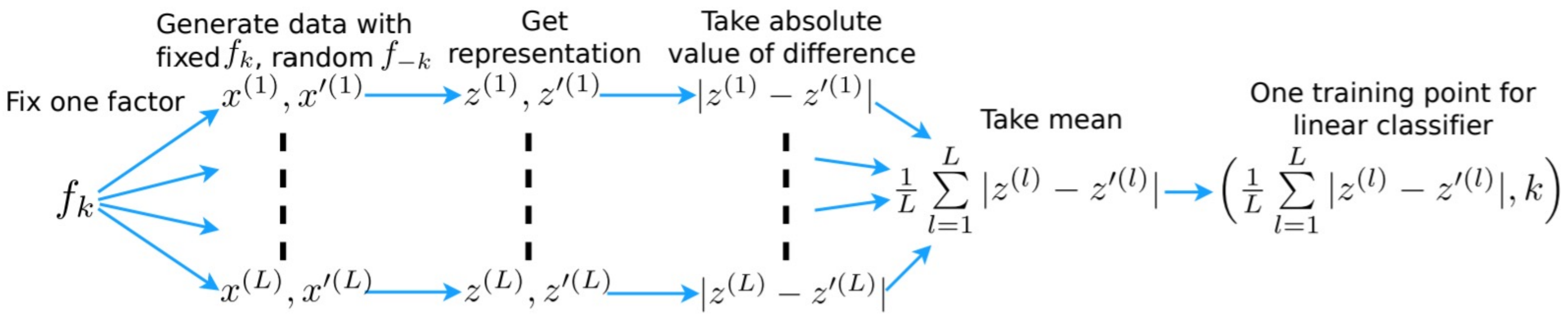
- Hilbert-Schmidt Independence Criterion (HSIC)
  → <u>Independence test</u> : $x\ and\ y\ independent? \rightarrow$ if $yes,\ 0$
  - $H(p,q) = \left\| E_{x\sim p,\, y\sim q}\left[\left(f(x)-\mu_x\right)\otimes\left(g(y)-\mu_y\right)\right]\right\|_{HS}^2$
    $= E_{x,x'\sim p,\, y,y'\sim q}[k(x,x')l(y,y')] + E_{x,x'\sim p}\,[k(x,x')]E_{y,y'\sim q}\,[l(y,y')] - 2E_{x,y}\left[E_{x'}[k(x,x')]E_{y'}[l(y,y')]\right]$

- Maximum Mean Discrepancy (MMD)
  → <u>Two sample test</u> : $\{x_i\}\sim p(x)\ and\ \{y_i\}\sim q(y) from\ \ the\ \ same\ \ dist? \rightarrow if\ yes,\ 0$
  - $M(p,q) = \max_{h\in H}\{E_{x\sim p}[h(x)] - E_{y\sim q}[h(y)]\ \ s.t.\ \|h\|_H \le 1\}$
    $= E_{x,x'\sim p,\, y,y'\sim q}\,[k(x,x') + k(y,y') - 2k(x,y')]$

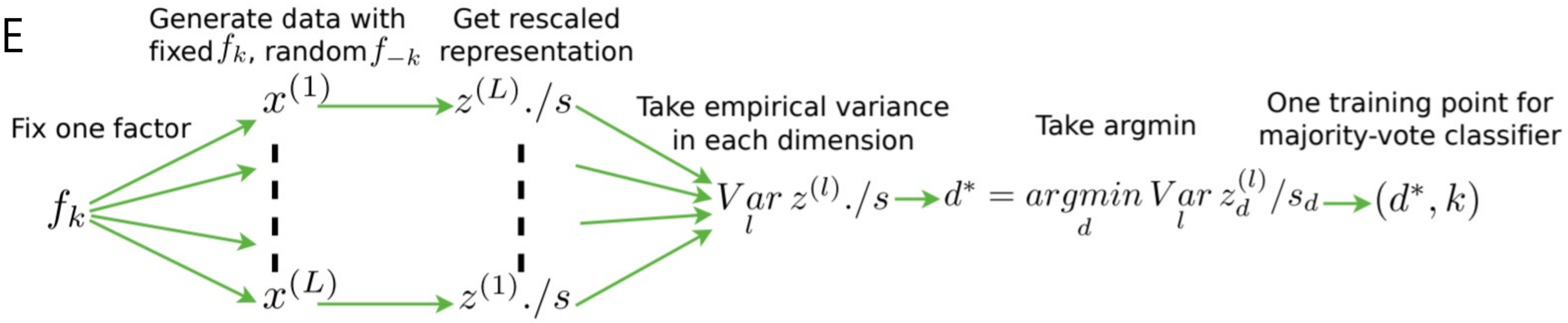
- Kernelized Stein Discrepancy (KSD)
  → <u>Goodness-of-fit test</u> : $\{x_i\}\sim q(x)\ from\ \ p(x)? \rightarrow if\ yes,\ 0$
  - $A_p f(x) \coloneqq \nabla_x \log p(x)\, f(x)^T + \nabla_x f(x) \quad \rightarrow \quad E_{x\sim p}\left[A_p f(x)\right] = 0\ \ if\ \ f\ \ is\ \ in\ \ stein\ \ class\ \ of\ \ p$
  - $K(q,p) = \max_{f\in H^d}\left\{E_{x\sim q}\left[trace\left(A_p f(x)\right)\right]\ \ s.t.\ \|f\|_{H^d} \le 1\right\}$
    $= \left\|E_{x\sim q}\left[A_p k(\cdot,x)\right]\right\|_{H^d}\ \ where\ \ f_{opt} = E_{x\sim q}\left[A_p k(\cdot,x)\right] / \left\|E_{x\sim q}\left[A_p k(\cdot,x)\right]\right\|_{H^d}$

Disentanglement measure

1. From Beta-VAE



2. From Factor-VAE



3. From Beta-TC-VAE

- $Mutual\ Information\ Gap\ (MIG) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H(v_k)}\left(I_n\left(z_{j^{(k)}}; v_k\right) - \max_{j \neq j^{(k)}} I_n(z_j; v_k)\right)$

  $where\ \ j^k = argmax_j I_n(z_j; v_k),\ \ H(v_k) = E_{p(v_k)}[-\log p(v_k)]\ \ and\ \ 0 \leq I(z_j; v_k) \leq H(v_k)$

Explicit distribution

1. Disentanglement
    1. Beta-VAE
    2. Factor-VAE
    3. Beta-TC-VAE
    4. HSIC-constrained-VAE

2. Gaussian Mixture Model
    1. Variational Deep Embedding (VaDE)
    2. Gaussian Mixture VAE (GM-VAE)

3. Dirichlet Mixture Model
    1. Stick Breaking VAE (SB-VAE)
    2. Dirichlet VAE (Dir-VAE)

4. Flow-based
    1. Normalizing Flows (NF)
    2. Inverse Autoregressive Flows (IAF)

Gaussian Mixture Model

1. Variational Deep Embedding (VaDE)

- Joint distribution : $p_\theta(x,z,c) = p_\theta(x|z)p_\theta(z|c)p(c)$

$$p_\theta(x|z) = Ber(x|\mu_x(z);\theta) \quad or \quad N(x|\mu_x(z),\sigma_x^2(z)I;\theta)$$
$$p(z|c) = N(z|\mu_z(c),\sigma_z^2(c)I)$$
$$p(c) = Cat(c|\pi)$$

- Variational distribution : $q_\phi(z,c|x) = q_\phi(z|x)q(c|x)$

$$q_\phi(z|x) = N\left(z\middle|\tilde{\mu}_z(x),\tilde{\sigma}_z{}^2(x)I;\phi\right)$$
$$q(c|x) \approx p(c|z) = p(c)p(z|c)/\sum_{c'=1}^{K} p(c')p(z|c')$$
$$(\because L_{ELBO}(X) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL\left(q_\phi(z|x)\|p(z)\right) - \cancel{KL(q(c|x)\|p(c|z))}])$$

- ELBO

$$L_{ELBO}(x) = E_{q_\phi(z,c|x)}\left[\log p(x,z,c) - \log q_\phi(z,c|x)\right]$$
$$= E_{q_\phi(z,c|x)}\left[\log p_\theta(x|z) + \log p(z|c) + \log p(c) - \log q_\phi(z|x) - \log q(c|x)\right]$$

Gaussian Mixture Model

2. Gaussian Mixture VAE (GM-VAE)

- Joint distribution : $p_\theta(x,z,w,c) = p_\theta(x|z)p_\theta(z|w,c)p(w)p(c)$

$$p_\theta(x|z) = Ber(x|\mu_x(z);\theta) \quad or \quad N(x|\mu_x(z),\sigma_x^2(z)I;\theta)$$
$$p_\theta(z|w,c) = N(z|\mu_z(w,c),\sigma_z^2(w,c)I;\theta)$$
$$p(w) = N(w|0,I)$$
$$p(c) = Cat(c|\pi)$$

- Variational distribution : $q_\phi(z,c,w|x) = q_\phi(z|x)q_\phi(w|x)q(c|z,w)$

$$q_\phi(z|x) = N\left(z\middle|\tilde{\mu}_z(x),\tilde{\sigma}_z{}^2(x)I;\phi\right)$$
$$q_\phi(w|x) = N\left(z\middle|\hat{\mu}_w(x),\hat{\sigma}_w{}^2(x)I;\phi\right)$$
$$q(c|z,w) \approx p_\theta(c|z,w) = p(c)p_\theta(z|w,c)/\sum_{c'=1}^{K} p(c')p_\theta(z|w,c')$$

- ELBO

$$\begin{aligned} L_{ELBO}(x) &= E_{q_\phi(Z,c,w|x)}[\log p_\theta(x,z,w,c) - \log q_\phi(z,c,w|x)] \\ &= E_{q_\phi(z|x)}[\log p_\theta(x|z)] - E_{q_\phi(w|x)p_\theta(c|Z,w)}\left[KL\left(q_\phi(z|x)\|p_\theta(z|w,c)\right)\right] \\ &\quad -KL\left(q_\phi(w|x)\|p(w)\right) - E_{q_\phi(Z|x)q_\phi(w|x)}[KL(p_\theta(c|z,w)\|p(c))] \end{aligned}$$

Explicit distribution

1. Disentanglement
   1. Beta-VAE
   2. Factor-VAE
   3. Beta-TC-VAE
   4. HSIC-constrained-VAE

2. Gaussian Mixture Model
   1. Variational Deep Embedding (VaDE)
   2. Gaussian Mixture VAE (GM-VAE)

3. Dirichlet Mixture Model
   1. Stick Breaking VAE (SB-VAE) → introduce stochastic width, resolve decoder weight collapsing
   2. Dirichlet VAE (Dir-VAE) → appropriate for multimodal posterior, resolve latent value collapsing

4. Flow-based
   1. Normalizing Flows (NF)
   2. Inverse Autoregressive Flows (IAF)

Dirichlet Mixture Model

1. Stick Breaking VAE (SB-VAE)

- Joint distribution : $p_\theta(x,\pi) = p_\theta(x|\pi)p(\pi) = p_\theta(x|\pi)p(v)$

$$p_\theta(x|\pi) = Ber(x|\mu_x(\pi);\theta) \quad or \quad N(x|\mu_x(\pi), \sigma_x^2(\pi)I;\theta)$$

$$\pi_k = \begin{cases} v_1 & if \;\; k=1 \\ v_k \prod_{j<k}(1-v_j) & if \;\; k>1 \end{cases} \quad where \;\; v_k \sim Beta\,(v_k|1,\alpha_0)$$

- Variational distribution : $q_\phi(\pi|x) = q_\phi(v|x)$

$$\pi_k = \begin{cases} v_1 & if \;\; k=1 \\ v_k \prod_{j<k}(1-v_j) & if \;\; k>1 \end{cases} \quad where \;\; v_k \sim Kumaraswamy\,(v_k|a(x),b(x);\phi) \;\; and \;\; v_K = 1(\approx Truncation)$$

$$*\ Reparameterization: \;\; v_k \approx \left(1-u^{\frac{1}{b(x)}}\right)^{\frac{1}{a(x)}} \quad where \;\; u \sim Uniform(0,1)$$

- ELBO

$$L_{ELBO}(X) = E_{q_\phi(v|x)}[\log p(x|\pi)] - \boxed{KL\big(q_\phi(v|x)\|p(v)\big)}$$

$$\boxed{\frac{a(x)-1}{a(x)}\left(-\gamma-\Psi\big(b(x)\big)-\frac{1}{b(x)}\right)+\log a(x)b(x)+\log B(1,\alpha_0)-\frac{b(x)-1}{b(x)}+(\alpha_0-1)b(x)\sum_{m=1}^{\infty}\frac{1}{m+a(x)b(x)}B\left(\frac{m}{a(x)},b(x)\right)}$$

Dirichlet Mixture Model

2. Dirichlet VAE (Dir-VAE)

- Joint distribution : $p_\theta(x,z) = p_\theta(x|z)p(z)$

$$p_\theta(x|z) = Ber(x|\mu_x(z);\theta) \quad or \quad N(x|\mu_x(z),\sigma_x^2(z)I;\theta)$$
$$p(z) = Dirichlet(\alpha) \ where \ \alpha = \left(1-\frac{1}{K},\dots,1-\frac{1}{K}\right)$$

- Variational distribution : $q_\phi(z|x)$

$$q_\phi(z|x) = Dirichlet\big(\mathrm{a}(x)\big) \ when \ z_k = \frac{v_k}{\sum_{k'} v_{k'}} \quad where \ v_k \sim Gamma(v_k|a(x),1)$$
$$* \ Reparameterization: \ v_k \approx \big(ua(x)\Gamma(\mathrm{a}(\mathrm{x}))\big)^{\frac{1}{a(x)}} \quad where \ u \sim Uniform(0,1)$$

- ELBO

$$L_{ELBO}(X) = E_{q_\phi(v|x)}[\log p(x|\pi)] - \boxed{KL\big(q_\phi(z|x)\|p(z)\big)}$$

$$\boxed{\sum_k [\log\Gamma(\alpha_k) - \log\Gamma(a(x)_k) + (a(x)_k - \alpha_k)\Psi(a(x)_k)]}$$

Explicit distribution

1. Disentanglement
   1. Beta-VAE
   2. Factor-VAE
   3. Beta-TC-VAE
   4. HSIC-constrained-VAE
2. Gaussian Mixture Model
   1. Variational Deep Embedding (VaDE)
   2. Gaussian Mixture VAE (GM-VAE)
3. Dirichlet Mixture Model
   1. Stick Breaking VAE (SB-VAE)
   2. Dirichlet VAE (Dir-VAE)

VAE-based

4. Flow-based
   1. Normalizing Flows (NF)
   2. Inverse Autoregressive Flows (IAF)

Restrictive하게 posterior의 form을 미리 정하지 않고, 간단한 distribution을 여러 차례 transform하는 방식을 택함으로써 ELBO를 tight하게 만듦.

Flow-based

1. Normalizing Flow (NF)
   - $z_K = f_K \circ f_{K-1} \circ \cdots \circ f_1(z_0)$
     (ex. $f_k(z_{k-1}) = z_{k-1} + u_k \cdot h_k\left(w_k^T z_{k-1} + b_k\right)$)
   - $\log q_K(z_K|x) = \log q_0(z_0|x) - \sum_{k=1}^{K} \log\left|\det \frac{\partial f_k}{\partial z_{k-1}}\right|$
     $= \log q_0(z_0|x) - \sum_{k=1}^{K} \log\left|1 + u_k^T \cdot h_k'\left(w_k^T z_{k-1} + b_k\right) w_k\right|$

2. Inverse Autoregressive Flow (IAF)
   - $z_0 = f_0(\epsilon) = \mu_0(x) + \sigma_0(x) \odot \epsilon \quad where \;\; p(\epsilon) = N(0, I_D)$
   - $z_k = f_k(z_{k-1}) = \mu_k(h(x), z_{k-1}) + \sigma_k(h(x), z_{k-1}) \odot z_{k-1} \quad where \;\; 1 \le k \le K$
   - $\log q_K(z_K|x) = \log p(\epsilon) - \sum_{k=0}^{K} \log\left|\det \frac{\partial f_k}{\partial z_{k-1}}\right|$
     $= \sum_{i=1}^{D}\left[-\frac{1}{2}\log 2\pi - \frac{1}{2}\epsilon_i^2 - \sum_{k=0}^{K} \log \sigma_{k,i}\right]$

Implicit distribution

1. GAN-based
    1. Adversarial Autoencoder (AAE)
    2. Info-GANs

2. Kernel-based
    1. Stein Variational Gradient Descent (SVGD)

GAN-based

1. Adversarial Autoencoder (AAE)

   1st step : $\max_{\theta,\phi} \mathrm{E}_{\mathrm{z}\sim \mathrm{q}_\phi(\mathrm{z}|\mathrm{x})}[\log p_\theta(x|z)]$

   2nd step : $\min_{\phi} \max_{\psi} E_{z\sim p(z)}\left[\log D_\psi(z)\right] + E_{z\sim q_\phi(z)}\left[\log(1-D_\psi(z))\right]$

   $(q_\phi(z) = E_{p(x)}[q(z|x)])$

2. Info-GANs

   - $\min_{G} \max_{D} E_{x\sim p(x)}\,[\log D(x)] + E_{z\sim p(z),\,c\sim p(c)}\left[\log(1-D\big(G(z,c)\big))\right] - \lambda \cdot I\big(c; G(z,c)\big)$
     $\approx \min_{G,q} \max_{D} E_{x\sim p(x)}\,[\log D(x)] + E_{z\sim p(z),\,c\sim p(c)}\left[\log(1-D\big(G(z,c)\big))\right] - \lambda \cdot [H(c) + E_{z\sim p(z),\,c\sim p(c), x\sim G(z,c)}[\log q(c|x)]]$

$$
\begin{aligned}
{}^{*}\; I\big(c; G(z,c)\big) &= H(c) - H(c|G(z,c)) \\
&= H(c) + E_{z\sim p(z),c\sim p(c),x\sim G(z,c)}\left[E_{c'\sim p(c'|x)}[\log p(c'|x)]\right] \\
&= H(c) + E_{z\sim p(z),c\sim p(c),x\sim G(z,c)}\left[KL(p(c'|x)\|q(c'|x)) + E_{c'\sim p(c'|x)}[\log q(c'|x)]\right] \\
&\geq H(c) + E_{z\sim p(z),c\sim p(c),x\sim G(z,c),c'\sim p(c'|x)}[\log q(c'|x)] \\
&= H(c) + E_{z\sim p(z),c\sim p(c),x\sim G(z,c)}[\log q(c|x)]
\end{aligned}
$$

Implicit distribution

1. GAN-based
    1. Adversarial Autoencoder (AAE)
    2. Info-GANs

2. Kernel-based
    1. Stein Variational Gradient Descent (SVGD)

Kernel-based

1. Stein Variational Gradient Descent (SVGD) : $T(z) = z + \epsilon \cdot f(z) \ \ (z \sim q \ \ and \ \ T(z) \sim q_{[T]})$

   ** Main theorem

$$\nabla_\epsilon KL\left(q_{[T]}(\cdot) \| p(\cdot)\right)\Big|_{\epsilon=0} = -E_{z\sim q}\left[trace\left(A_p f(z)\right)\right]$$

   ** Main algorithm

$$T(z) = z + \epsilon \cdot E_{z'\sim q}\left[A_p k(z, z')\right] \approx z + \frac{1}{n}\sum_{j=1}^{n}\left[\nabla_{z^j}\log \mathrm{p}\left(z^j\right)k\left(z, z^j\right) + \nabla_{z^j} k\left(z, z^j\right)\right]$$

   ** Interpretation as Functional Gradient Descent (FGD)

$$\lim_{\epsilon \to 0} \frac{L(f + \tilde{\epsilon} \cdot g) - L(f)}{\tilde{\epsilon}} \coloneqq \left\langle \nabla_f L(f), g \right\rangle_{H_{k(\cdot,\cdot)}}$$

$$L(f) \coloneqq KL\left(q_{[T]}(\cdot) \| \, p(\cdot)\right) \quad \Longrightarrow \quad \nabla_f L(f)\Big|_{f=0}(z) = -E_{z'\sim q}\left[A_p k(z, z')\right]$$

$$\therefore T(z) = z + \epsilon \cdot f(z) = z - \epsilon \cdot \nabla_f L(f)\Big|_{f=0}(z)$$

Clustering

1. Deep Embedded Clustering (DEC)
2. Set Transformer (ST)
3. Deep Amortized Clustering (DAC)

Clustering

1. Deep Embedded Clustering (DEC)

- Using student t's distribution, measure the similarity between embedded point z_i and centroid μ_j

$$q_{ij} = \frac{\left(1 + \left\|z_i - \mu_j\right\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1 + \left\|z_i - \mu_{j'}\right\|^2/\alpha\right)^{-\frac{\alpha+1}{2}}} \ : \ Probability\ of\ assigning\ sample\ i\ to\ cluster\ j$$

- Set the target distribution

$$p_{ij} = \frac{q_{ij}^2/\sum_i q_{ij}}{\sum_{j'} q_{ij'}^2/\sum_i q_{ij'}}$$

1. Improve cluster purity
2. Put more emphasis on data points assigned with high confidence
3. Prevent large clusters from distortion by normalization

- Update the embedding function $f_\theta: X \to Z$ and the cluster centroid μ with SGD

$$Loss = KL(P||Q) = \sum_i \sum_j p_{ij} \log p_{ij}/q_{ij}$$

Clustering

Randomly Initialized Learnable Parameters

2. Set Transformer (ST)
   - $X \in R^{n\times d}$, $Y \in R^{n\times d}$, $I \in R^{m\times d}$, $S \in R^{k\times d}$, $Z = R^{n\times d}$
   - Operations & Blocks
     - rFF : $row-wise\ feedforward\ layer$
     - $MAB(X,Y) = LayerNorm\big(H + rFF(H)\big) \in R^{n\times d}$ $where\ H = LayerNorm\big(X + MultiHead(X,Y,Y)\big)$
     - $SAB(X) = MAB(X,X) \in R^{n\times d}$
     - $ISAB_m(X) = MAB(X,H) \in R^{n\times d} where\ H = MAB(I,X) \in R^{m\times d}$
     - $PMA_k(Z) = MAB\big(S, rFF(Z)\big) \in R^{k\times d}$

   - $Encoder(X) = SAB\big(SAB(X)\big) \in R^{n\times d}$ or $ISAB_m\big(ISAB_m(X)\big) \in R^{n\times d}$
   - $Decoder(Z) = rFF\left(SAB\big(PMA_k(Z)\big)\right) \in R^{k\times d}$

** Main theorem
Set Transformer is a universal approximator of permutation invariant functions.

** Amortized Clustering with Mixture of Gaussians

$$\max E_X\left[\sum_{i=1}^{|X|}\log\sum_{j=1}^{k}\pi_j(X)N\left(x_i;\mu_j(X),diag\left(\sigma_j^2(X)\right)\right)\right]\ where\ \{\pi_j(X),\mu_j(X),\log\sigma_j^2(X)\}_{j=1}^{k}\ is\ output$$

Clustering

3. Deep Amortized Clustering (DAC)
   - ST + Filtering

1. Minimum Loss Filtering

$$L(x, y, m, \theta) = \min_{j \in \{1,\dots,k_X\}} \left(\frac{1}{n_X} \sum_{i=1}^{n_X} BCE\big(m_i, I(y_i = j)\big) - \frac{1}{n_{X|j}} \sum_{i|y_i=j} \log p(x_i; \theta) \right)$$

$$\begin{array}{rll} \textbf{encode data:} & H_X = \mathrm{ISAB}_L(X), & \\ \textbf{decode cluster:} & H_\theta = \mathrm{PMA}_1(H_X), & \theta = \mathrm{rFF}(H_\theta), \\ \textbf{decode mask:} & H_{\mathfrak{m}} = \mathrm{ISAB}_{L'}(\mathrm{MAB}(H_X, H_\theta)), & \mathfrak{m} = \mathrm{sigmoid}(\mathrm{rFF}(H_{\mathfrak{m}})) \end{array}$$

2. Anchored Filtering

$$L(x, y, a, m, \theta) = \frac{1}{n_X} \sum_{i=1}^{n_X} BCE\big(m_i, I(y_i = j_a)\big) - \frac{1}{n_{X|j_a}} \sum_{i|y_i=j_a} \log p(x_i; \theta)$$

$$\begin{array}{rll} \textbf{encode data:} & H_X = \mathrm{ISAB}_L(X), & H_{X|a} = \mathrm{MAB}(H_X, h_a), \\ \textbf{decode cluster:} & H_\theta = \mathrm{PMA}_1(H_{X|a}), & \theta = \mathrm{rFF}(H_\theta), \\ \textbf{decode mask:} & H_{\mathfrak{m}} = \mathrm{ISAB}_{L'}(\mathrm{MAB}(H_{X|a}, H_\theta)), & \mathfrak{m} = \mathrm{sigmoid}(\mathrm{rFF}(H_{\mathfrak{m}})) \end{array}$$

Evaluation Metrics in BNN

- Fidelity of posterior approximation
  - $\widehat{y_n}^{high}$ and $\widehat{y_n}^{low}$ are 97.5% and 2.5% percentile, respectively

  1. Average Marginal Log-likelihood (higher the better) : $E_{(x_n,y_n)\sim D}\left[E_{q(w)}[p(y_n|x_n,w)]\right]$
  2. Predictive RMSE (lower the better) : $\sqrt{\frac{1}{N}\sum_{n=1}^{N}\left\|y_n - E_{q(w)}[f x_n, w)]\right\|_2^2}$
  3. Prediction Interval Coverage Probability (PICP) (higher the better) : $\frac{1}{N}\sum_{n=1}^{N} I\left[y_n \le \widehat{y_n}^{high}\right] \cdot I\left[y_n \ge \widehat{y_n}^{low}\right]$
  4. Mean Prediction Interval Width (MPIW) (lower the better) : $\frac{1}{N}\sum_{n=1}^{N}(\widehat{y_n}^{high} - \widehat{y_n}^{low})$

- Uncertainty Calibration

  1. Reliability diagram : $p(correct\ |confidence = \rho)\ \ \forall\rho \in [0,1]$
  2. Expected Calibration Error (ECE) (lower the better) : $ECE = E_{confidence}[|p(correct|confidence) - confidence)|]$