电子科技大学
University of Electronic Science and Technology of China

Information Theory in Deep Neural Networks

Wei Han

DM LESS IS MORE

Data Mining Lab,
Big Data Research Center, UESTC
Email： weihan@std.uestc.edu.cn

Outline

- Preliminary of Information Theory
- Information between Data and Representation
- Information Bottleneck & Mutual Information
- Disentangled Representation via Mutual Information
- Generalization Bound from Mutual Information

Preliminary

1.1. Entropy

- Uncertainty

- Shannon Entropy

$$H(X) = -\sum_{x \in \chi} p(x) log p(x)$$

1.1. Entropy

- Uncertainty

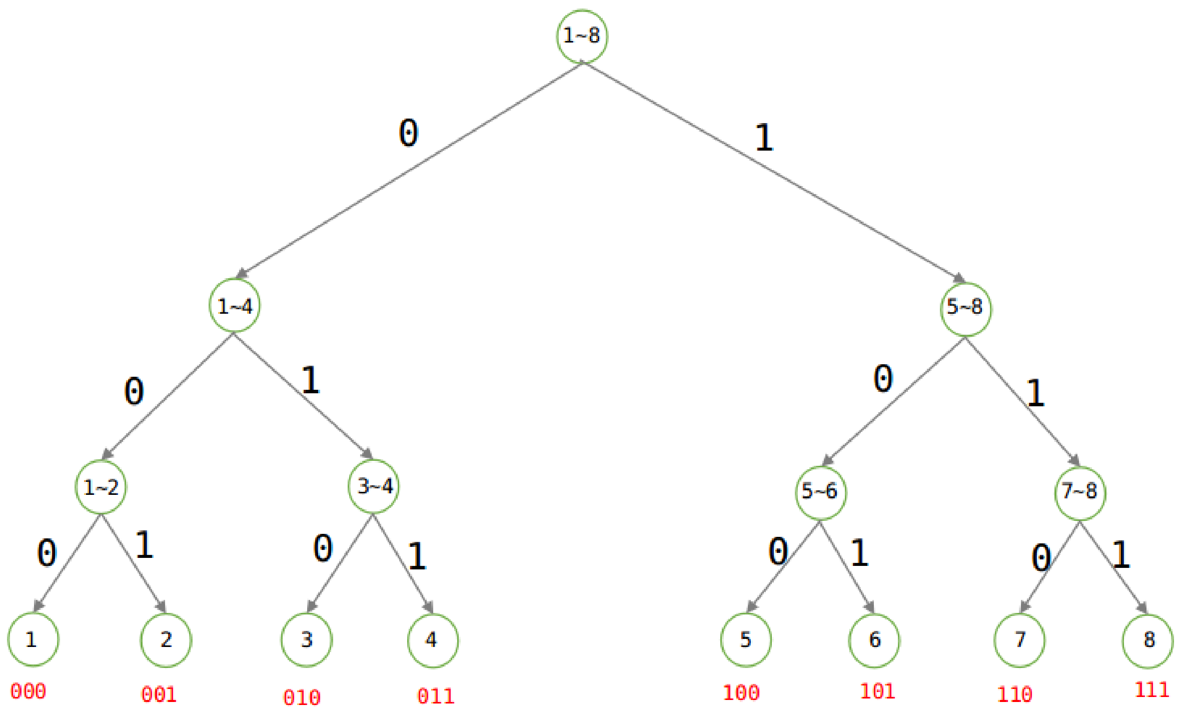
$$H(X) = -\frac{1}{2} \times log\frac{1}{2} - \frac{1}{2} \times log\frac{1}{2} = 1$$

$$P\{Y = 1, 2, \ldots, 6\} = \frac{1}{6} \quad H(Y) = 6 \times \left(-\frac{1}{6} log\frac{1}{6}\right) = log6$$

$$H(X) = 1 = log2 < H(Y) = log6$$

- Minimum MCL

$$H(x) = -\sum_{x \in X} p(x) \log p(x) = E_{x \sim P}[-\log p(X)]$$



1.3. Cross Entropy

- Coding P with Q

$$H(P,Q) = E_{x \sim P}[-\log Q(X)] = -\sum_{x \in X} \underbrace{p(x) \underbrace{\log q(x)}_{\text{Coding}}}_{\text{Minimize MCL}}$$

1.4. KL Divergence

- Distribution Gap

$$KL(P||Q) = \sum P(x)\log\frac{P(x)}{Q(x)} = \sum_{x\in X} p(x)\log p(x) - \sum_{x\in X} p(x)\log q(x)$$

$$= \underline{-H(P)} + H(P,Q)$$

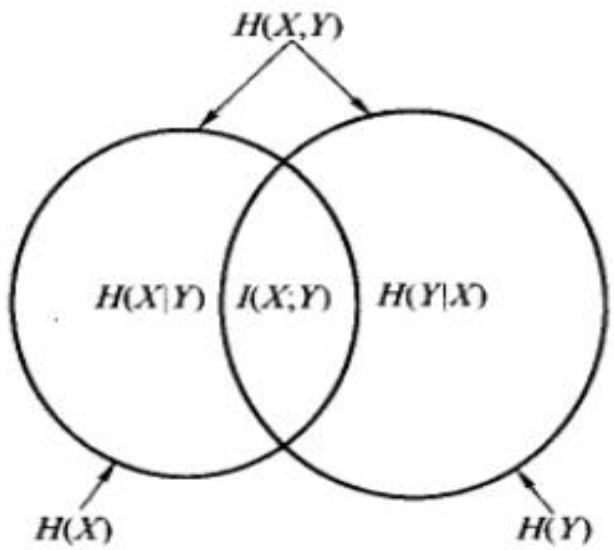
Data are given

1.5. Mutual Information

- Dependence

$$\mathrm{I}(X;Y) = \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

$$I(X;Y) = H(X) - H(X|Y).$$



Information between
Data and Representation

2.1. Minimum Entropy in Data

数据挖掘实验室
Data Mining Lab

- Information Entropy

$$\mathcal{H}_c = -\sum_{c \in \text{汉字}} p_c \log p_c$$

Chinese character *9.65bit* vs. English character *4.03bit*

From characters to words

$$-\sum_{w \in \text{语料}} \log p_w$$

2.1. Minimum Entropy in Data

数据挖掘实验室
Data Mining Lab

- Information Entropy

$$-\sum_{w\in 语料}\log p_w < -\sum_{c\in 语料}\log p_c \longrightarrow -\sum_{w\in 词表}N_w\log p_w < -\sum_{c\in 字表}N_c\log p_c$$

$$-\sum_{c\in 字表}\frac{N_c}{总字数}\log p_c = -\sum_{c\in 字表}p_c\log p_c$$

$$\begin{aligned}\mathcal{L} &= -\sum_{w\in 词表}\frac{N_w}{总字数}\log p_w\\ &=\left(-\sum_{w\in 词表}\frac{N_w}{总词数}\log p_w\right)\div\left(\frac{总字数}{总词数}\right)\\ &=\left(-\sum_{w\in 词表}\frac{N_w}{总词数}\log p_w\right)\div\left(\frac{\sum\limits_{w\in 词表}N_w l_w}{总词数}\right)\\ &=\left(-\sum_{w\in 词表}\frac{N_w}{总词数}\log p_w\right)\div\left(\sum_{w\in 词表}\frac{N_w}{总词数}l_w\right)\\ &=\frac{\mathcal{H}}{l}\end{aligned}$$

2.1. Minimum Entropy in Data

- Localize / Segment

$$\mathcal{H} = -\sum_{w\in\text{词表}} p_w \log p_w, \quad l = \sum_{w\in\text{词表}} p_w l_w \qquad \mathcal{L} = \frac{\mathcal{H}}{l} = \frac{-\sum_i p_i \log p_i}{\sum_i p_i l_i}$$

To combine character *a* and *b* as a word

$$\tilde{p}_{ab} = \frac{N_{ab}}{\tilde{N}} = \frac{p_{ab}}{1-p_{ab}}$$

$$\tilde{p}_a = \frac{\tilde{N}_a}{\tilde{N}} = \frac{p_a - p_{ab}}{1-p_{ab}}, \tilde{p}_b = \frac{\tilde{N}_b}{\tilde{N}} = \frac{p_b - p_{ab}}{1-p_{ab}}$$

$$\tilde{p}_i = \frac{N_i}{\tilde{N}} = \frac{p_i}{1-p_{ab}}, (i \neq a, b)$$

2.1. Minimum Entropy in Data

- Minimize Entropy

$$\tilde{\mathcal{H}} = -\frac{1}{1-p_{ab}}\left\{p_{ab}\log\left(\frac{p_{ab}}{1-p_{ab}}\right)+\right.$$

$$\left.\sum_{i=a,b}(p_i - p_{ab})\log\left(\frac{p_i - p_{ab}}{1-p_{ab}}\right) + \underline{\sum_{i\neq a,b} p_i \log\left(\frac{p_i}{1-p_{ab}}\right)}\right\}$$

$$=\frac{1}{1-p_{ab}}(\mathcal{H} - \underline{\mathcal{F}_{ab}})$$

$$\mathcal{F}_{ab} = p_{ab}\log\frac{p_{ab}}{p_a p_b} - (1-p_{ab})\log(1-p_{ab})$$

$$+\sum_{i=a,b}(p_i - p_{ab})\log\left(1-\frac{p_{ab}}{p_i}\right)$$

2.1. Minimum Entropy in Data

- Minimize Entropy

$$\tilde{l} = \frac{p_{ab}}{1-p_{ab}}(l_a + l_b) + \sum_{i=a,b} \frac{p_i - p_{ab}}{1-p_{ab}} l_i + \sum_{i \neq a,b} \frac{p_i}{1-p_{ab}} l_i$$
$$= \frac{l}{1-p_{ab}}$$

After knowing the change, the change of *L*:

$$\frac{\tilde{\mathcal{H}}}{\tilde{l}} - \frac{\mathcal{H}}{l} = -\frac{\mathcal{F}_{ab}}{l} \uparrow \quad \uparrow$$

2.1. Minimum Entropy in Data

- Minimize Entropy

$$\mathcal{F}_{ab} \approx F^*_{ab} = p_{ab}\left(\ln \frac{p_{ab}}{p_a p_b} - 1\right)$$

$$\mathcal{F}_{\boldsymbol{a}} \approx \mathcal{F}^*_{\boldsymbol{a}} = p_{\boldsymbol{a}}\left(\ln \frac{p_{\boldsymbol{a}}}{\prod_{i\in \boldsymbol{a}} p_i} - 1\right)$$

$$PMI(a,b) = ln\frac{p_{ab}}{p_a p_b} > 1$$

Segment

$$\updownarrow$$

$$ln\frac{p_{ab}}{p_a p_b} < min\_pmi$$

Split

2.2. Minimum Entropy in Representation

- Library Model

$$S = \sum_{i,j} p(i)\, d(i)$$

2.2. Minimum Entropy in Representation

- What if more than one book?

$$S = \sum_{i,j} p(i)p(j|i)[d(i) + d(i,j)] = \sum_{i,j} p(i,j)[d(i) + d(i,j)]$$

- book2vec

$$S = \min_{v} \sum_{i,j} p(i)p(j|i) \underline{\left[\|v_i\| + \|v_i - v_j\|\right]} = \sum_{i,j} p(i,j) \left[\|v_i\| + \|v_i - v_j\|\right]$$

$$\text{s.t.} \quad \forall i,j,\ \|v_i - v_j\| \geq d_{\min}$$

$$S = \sum_{i,j} p(i)p(j|i)f(v_i, v_j) = \sum_{i,j} p(i,j)\underline{f(v_i, v_j)}$$

2.2. Minimum Entropy in Representation

数据挖掘实验室
Data Mining Lab

- What if more than one book?

$$S = \sum_{i,j} p(i)p(j|i)f(v_i, v_j) = \sum_{i,j} p(i,j)f(v_i, v_j)$$

- word2vec – skip gram

$$f(v_i, v_j) = -\log \frac{e^{-\|v_i - v_j\|^2}}{Z_i}, \quad Z_i = \sum_j e^{-\|v_i - v_j\|^2}$$

$$f(v_i, v_j) = -\log \frac{e^{\langle v_i, v_j \rangle}}{Z_i}, \quad Z_i = \sum_j e^{\langle v_i, v_j \rangle}$$

$$S = \sum_{i,j} p(i)p(j|i)f(v_i, v_j) = -\sum_{i,j} p(i)p(j|i) \log q(j|i)$$

- (t-)SNE

Let $P(i) = c$

$$p(\boldsymbol{x}_j|\boldsymbol{x}_i) = \frac{e^{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2}}{\sum_j^{j\neq i} e^{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2}}$$

$$S = -\sum_{i,j}^{i\neq j} p(\boldsymbol{x}_j|\boldsymbol{x}_i)\log q(j|i), \quad q(j|i) = \frac{e^{-\|\boldsymbol{v}_i-\boldsymbol{v}_j\|^2}}{\sum_j^{j\neq i} e^{-\|\boldsymbol{v}_i-\boldsymbol{v}_j\|^2}}$$

- Generative model from Z to X

$$p(X) \sim \{X_1, \dots, X_n\}$$

$$p(x) = \int p(x|z)p(z)dz, \quad p(x,z) = \underline{p(x|z)p(z)}$$

$$KL\Big(p(x,z)\Big\|q(x,z)\Big) = \iint p(x,z)\ln\frac{p(x,z)}{q(x,z)}dzdx$$

$$\begin{aligned} KL\Big(p(x,z)\Big\|q(x,z)\Big) &= \int p(x)\left[\int p(z|x)\ln\frac{p(x,z)}{q(x,z)}dz\right]dx \\ &= \mathbb{E}_{x\sim p(x)}\left[\int p(z|x)\ln\frac{p(z|x)p(x)}{q(x,z)}dz\right] \end{aligned}$$

- Joint Probability Matching

$$KL\Big(p(x,z)\Big\|q(x,z)\Big) = \mathbb{E}_{x\sim p(x)}\left[\int p(z|x)\ln\frac{p(z|x)\boxed{p(x)}}{q(x,z)}dz\right]$$

$$\ln\frac{p(z|x)p(x)}{q(x,z)} = \ln\frac{p(z|x)}{q(x,z)} + \ln p(x)$$

$$\mathbb{E}_{x\sim p(x)}\left[\int p(z|x)\ln p(x)dz\right] = \mathbb{E}_{x\sim p(x)}\left[\ln p(x)\int p(z|x)dz\right]$$
$$= \mathbb{E}_{x\sim p(x)}[\ln p(x)]$$

$$\mathcal{L} = KL\Big(p(x,z)\Big\|q(x,z)\Big) - \text{常数} = \mathbb{E}_{x\sim p(x)}\left[\int p(z|x)\ln\frac{p(z|x)}{q(x,z)}dz\right]$$

2.3. VAE

数据挖掘实验室
Data Mining Lab

- Losses

$$
\begin{aligned}
\mathcal{L} &= \mathbb{E}_{x\sim p(x)}\left[\int p(z|x)\ln\frac{p(z|x)}{q(x|z)q(z)}dz\right] \\
&= \mathbb{E}_{x\sim p(x)}\left[-\int p(z|x)\ln q(x|z)dz + \int p(z|x)\ln\frac{p(z|x)}{q(z)}dz\right] \\
&= \mathbb{E}_{x\sim p(x)}\left[\mathbb{E}_{z\sim p(z|x)}\big[-\ln q(x|z)\big] + \mathbb{E}_{z\sim p(z|x)}\left[\ln\frac{p(z|x)}{q(z)}\right]\right] \\
&= \mathbb{E}_{x\sim p(x)}\left[\mathbb{E}_{z\sim p(z|x)}\big[\underline{-\ln q(x|z)}\big] + KL\Big(p(z|x)\Big\|\underline{q(z)}\Big)\right]
\end{aligned}
$$

Uniform Gaussian as a Prior

Loss for Generation

2.3. VAE

- Components

$$\mathcal{L} = \mathbb{E}_{x\sim p(x)}\left[\mathbb{E}_{z\sim p(z|x)}\left[-\ln \underbrace{q(x|z)}_{\text{Bernoulli / Gaussian}}\right] + KL\left(\underline{p(z|x)}\middle\|\underbrace{q(z)}_{\text{Standard Gaussian}}\right)\right]$$

$$\hat{p}(z|x) = q(z|x) = \frac{q(x|z)q(z)}{q(x)} = \frac{q(x|z)q(z)}{\underline{\int q(x|z)q(z)dz}}$$

2.3. VAE

- $P(z|x)$

$$p(z|x) = \frac{1}{\prod_{k=1}^{d} \sqrt{2\pi\sigma_{(k)}^2(x)}} \exp\left(-\frac{1}{2}\left\|\frac{z-\mu(x)}{\sigma(x)}\right\|^2\right)$$

$$KL\Big(p(z|x)\Big\|q(z)\Big) = \frac{1}{2}\sum_{k=1}^{d}\Big(\underline{\mu_{(k)}^2(x)} + \underline{\sigma_{(k)}^2(x)} - \ln\sigma_{(k)}^2(x) - 1\Big)$$

- $P(x|z)$ with Bernoulli distribution

$$p(\xi) = \begin{cases} \rho, \ \xi = 1; \\ 1 - \rho, \ \xi = 0 \end{cases}$$

$$q(x|z) = \prod_{k=1}^{D} \Big(\rho_{(k)}(z)\Big)^{x(k)} \Big(1 - \rho_{(k)}(z)\Big)^{1-x(k)}$$

$$-\ln q(x|z) = \sum_{k=1}^{D} \underbrace{\Big[- x_{(k)} \ln \rho_{(k)}(z) - (1 - x_{(k)}) \ln \Big(1 - \rho_{(k)}(z)\Big)\Big]}_{\text{Cross Entropy Loss}}$$

- *P(x|z)* with Gaussian distribution

$$q(x|z) = \frac{1}{\prod_{k=1}^{D} \sqrt{2\pi\sigma_{(k)}^2(z)}} \exp\left(-\frac{1}{2}\left\|\frac{x-\mu(z)}{\sigma(z)}\right\|^2\right)$$

$$-\ln q(x|z) = \frac{1}{2}\left\|\frac{x-\mu_k(z)}{\sigma(z)}\right\|^2 + \frac{D}{2}\ln 2\pi + \frac{1}{2}\sum_{k=1}^{D}\ln\sigma_{(k)}^2(z)$$

$$-\ln q(x|z) \sim \frac{1}{2\sigma^2}\underbrace{\left\|x-\mu_k(z)\right\|^2}_{\text{MSE}}$$

Information Bottleneck & Mutual Information

3.1. Information Bottleneck

- Information Extraction

$$\max[I(Z;Y) - \beta I(X;Z)]$$

- Loss

$I(X;Z)$

$$\underline{\iint p(z|x)\tilde{p}(x)\log\frac{p(z|x)}{p(z)}dxdz}$$

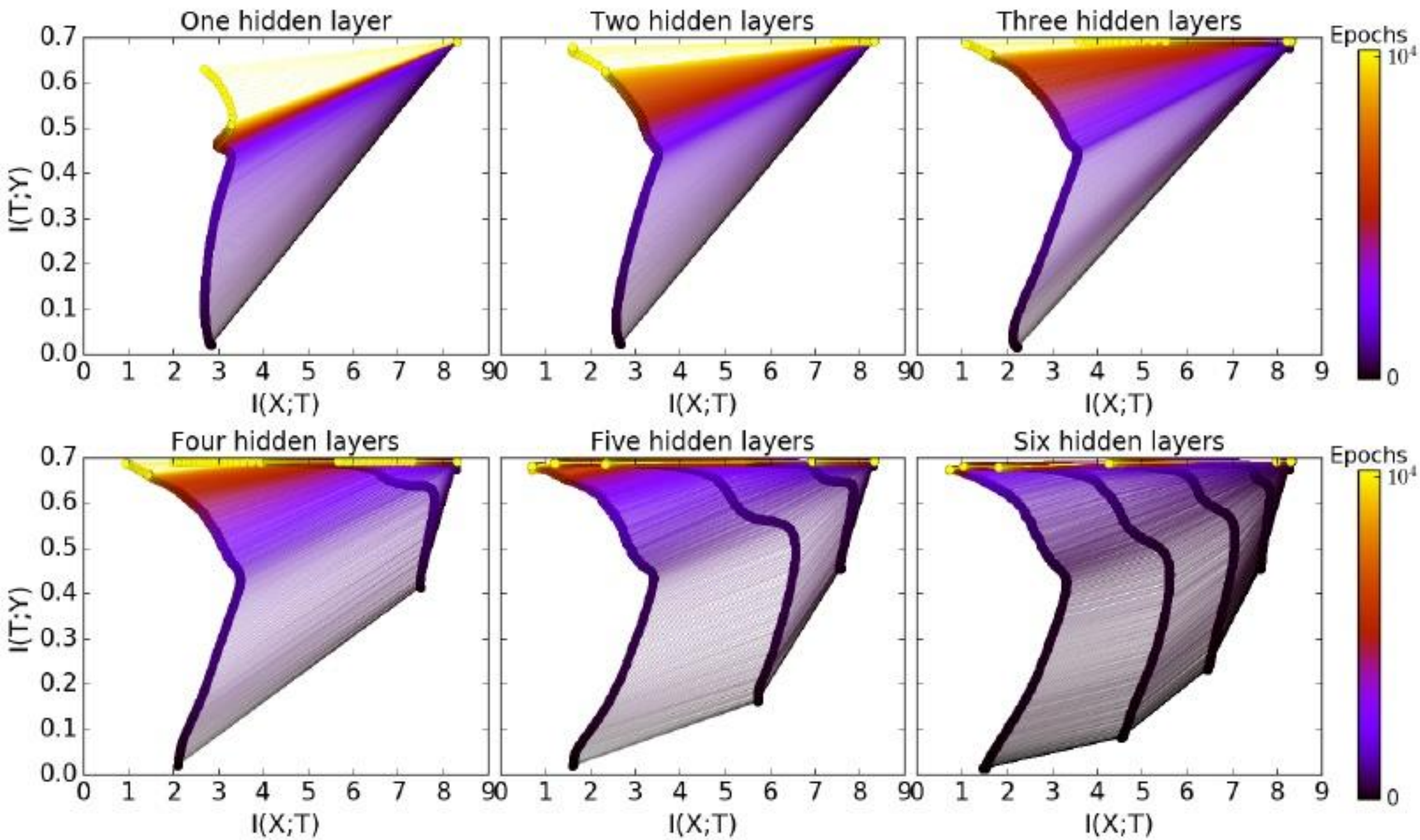
$$KL\left(p(z|x)\middle\|\underline{q(z)}\right)\Big]$$

Standard Gaussian

$$\begin{aligned}
&= \iint p(z|x)\tilde{p}(x)\log\frac{p(z|x)}{q(z)}\frac{q(z)}{p(z)}dxdz \\
&= \iint p(z|x)\tilde{p}(x)\log\frac{p(z|x)}{q(z)} + \iint p(z|x)\tilde{p}(x)\log\frac{q(z)}{p(z)}dxdz \\
&= \iint p(z|x)\tilde{p}(x)\log\frac{p(z|x)}{q(z)} + \int p(z)\log\frac{q(z)}{p(z)}dz \\
&= \iint p(z|x)\tilde{p}(x)\log\frac{p(z|x)}{q(z)} - \int p(z)\log\frac{p(z)}{q(z)}dz \\
&= \int \tilde{p}(x)KL\big(p(z|x)\big\|q(z)\big)dx - \underline{KL\big(\underline{p(z)}\big\|q(z)\big)} \\
&< \int \tilde{p}(x)KL\big(p(z|x)\big\|q(z)\big)dx
\end{aligned}$$

>0

- Two phase of learning



3.2. Deep InfoMax

- AutoEncoder – Keep key information

$$p(z|x) = \max_{p(z|x)} I(X, Z)$$

数据挖掘实验室
Data Mining Lab

- Maximum Mutual Information

$$I(X,Z) = \iint p(z|x)\tilde{p}(x) \log \frac{p(z|x)}{p(z)} dxdz$$

Given data distribution

$$p(z) = \int p(z|x)\tilde{p}(x)dx$$

$$p(z|x) = \max_{p(z|x)} I(X,Z)$$

数据挖掘实验室
Data Mining Lab

- Prior Distribution

$$KL(p(z)\|q(z)) = \int p(z)\log\frac{p(z)}{q(z)}dz$$

$$\begin{aligned} p(z|x) &= \min_{p(z|x)} \{-I(X,Z) + \lambda KL(p(z)\|q(z))\} \\ &= \min_{p(z|x)} \left\{ -\iint \underline{p(z|x)\tilde{p}(x)} \log\frac{p(z|x)}{p(z)} \underline{dxdz} + \lambda \int p(z)\log\frac{p(z)}{q(z)}dz \right\} \\ &= \min_{p(z|x)} \left\{ \iint \underline{p(z|x)\tilde{p}(x)} \left[-(1+\lambda)\log\frac{p(z|x)}{p(z)} + \lambda\log\frac{p(z|x)}{q(z)} \right] dxdz \right\} \\ &= \min_{p(z|x)} \left\{ -\beta \cdot \underline{I(X,Z)} + \gamma \cdot \underline{\mathbb{E}_{x\sim\tilde{p}(x)}[KL(p(z|x)\|q(z))]} \right\} \end{aligned}$$

3.2. Deep InfoMax

- Mutual Information

$$I(X,Z) = \iint p(z|x)\tilde{p}(x)\log\frac{p(z|x)\tilde{p}(x)}{p(z)\tilde{p}(x)}dxdz$$
$$= KL(p(z|x)\tilde{p}(x) \| p(z)\tilde{p}(x))$$

- From f-GAN

$$D_f(P\|Q) = \max_T \left(\mathbb{E}_{x\sim p(x)}[T(x)] - \mathbb{E}_{x\sim q(x)}[g(T(x))]\right)$$

$$JS\left(p(z|x)\tilde{p}(x), p(z)\tilde{p}(x)\right)$$
$$= \max_T \left(\mathbb{E}_{x\sim p(z|x)\tilde{p}(x)}[\log\sigma(T(x,z))] + \mathbb{E}_{x\sim p(z)\tilde{p}(x)}[\log(1-\sigma(T(x,z))]\right)$$

3.2. Deep InfoMax

- Deep InfoMax

$$p(z|x) = \min_{p(z|x)} \left\{-\beta \cdot JS\Big(p(z|x)\tilde{p}(x), p(z)\tilde{p}(x)\Big) + \gamma \cdot \mathbb{E}_{x\sim\tilde{p}(x)}[KL(p(z|x)\|q(z))]\right\}$$

$$\begin{aligned} &p(z|x), T(x,z) \\ = &\min_{p(z|x),T(x,z)} \Big\{ -\beta \cdot \Big(\mathbb{E}_{x\sim p(z|x)\tilde{p}(x)}[\log \sigma(\underline{T(x,z)})] + \mathbb{E}_{x\sim p(z)\tilde{p}(x)}[\log(1-\sigma(\underline{T(x,z)})]\Big) \\ &\qquad + \gamma \cdot \underline{\mathbb{E}_{x\sim\tilde{p}(x)}[KL(p(z|x)\|q(z))]} \Big\} \end{aligned}$$

Discriminator

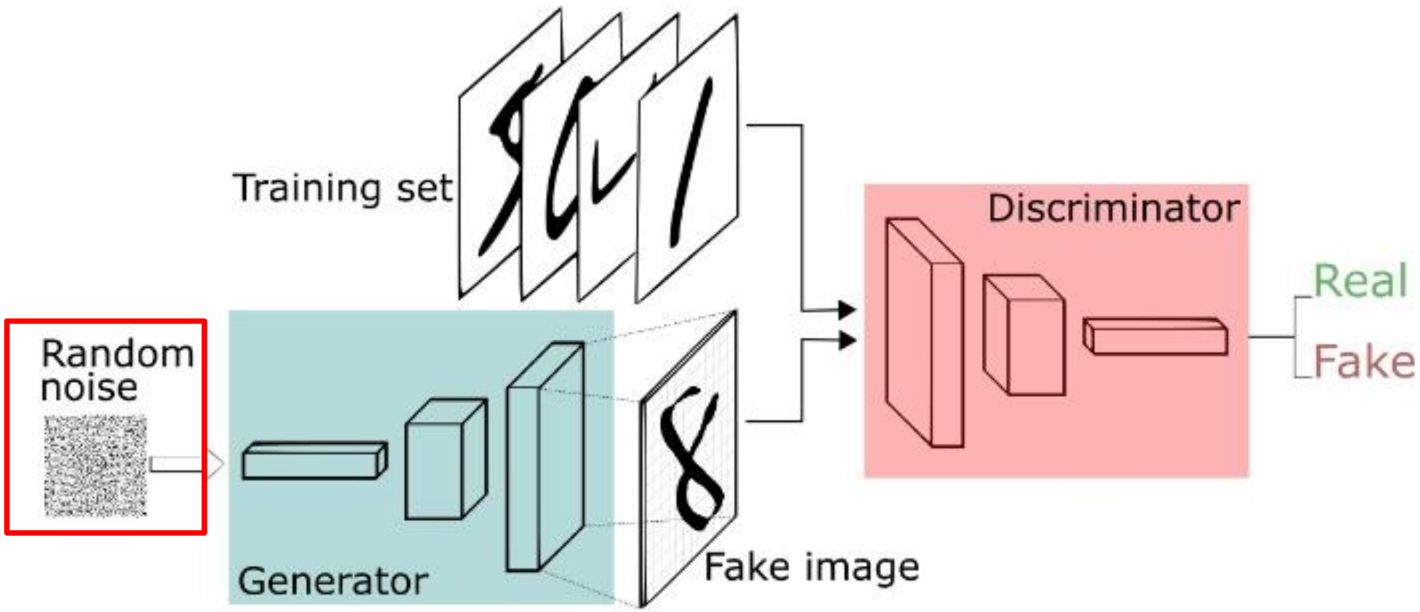
$$KL\Big(p(z|x)\Big\|q(z)\Big) = \frac{1}{2}\sum_{k=1}^{d}\Big(\mu_{(k)}^2(x) + \sigma_{(k)}^2(x) - \ln \sigma_{(k)}^2(x) - 1\Big)$$

Disentangled Representation

4.1. InfoGAN

- GAN

$$\min_{G}\max_{D} V(D,G) = E_{p(\mathbf{x})}[\log(D(\mathbf{x}))] + E_{p(\mathbf{z})}[\log(1 - D(G(\mathbf{z})))]$$



4.1. InfoGAN

数据挖掘实验室

Data Mining Lab

- InfoGAN

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$

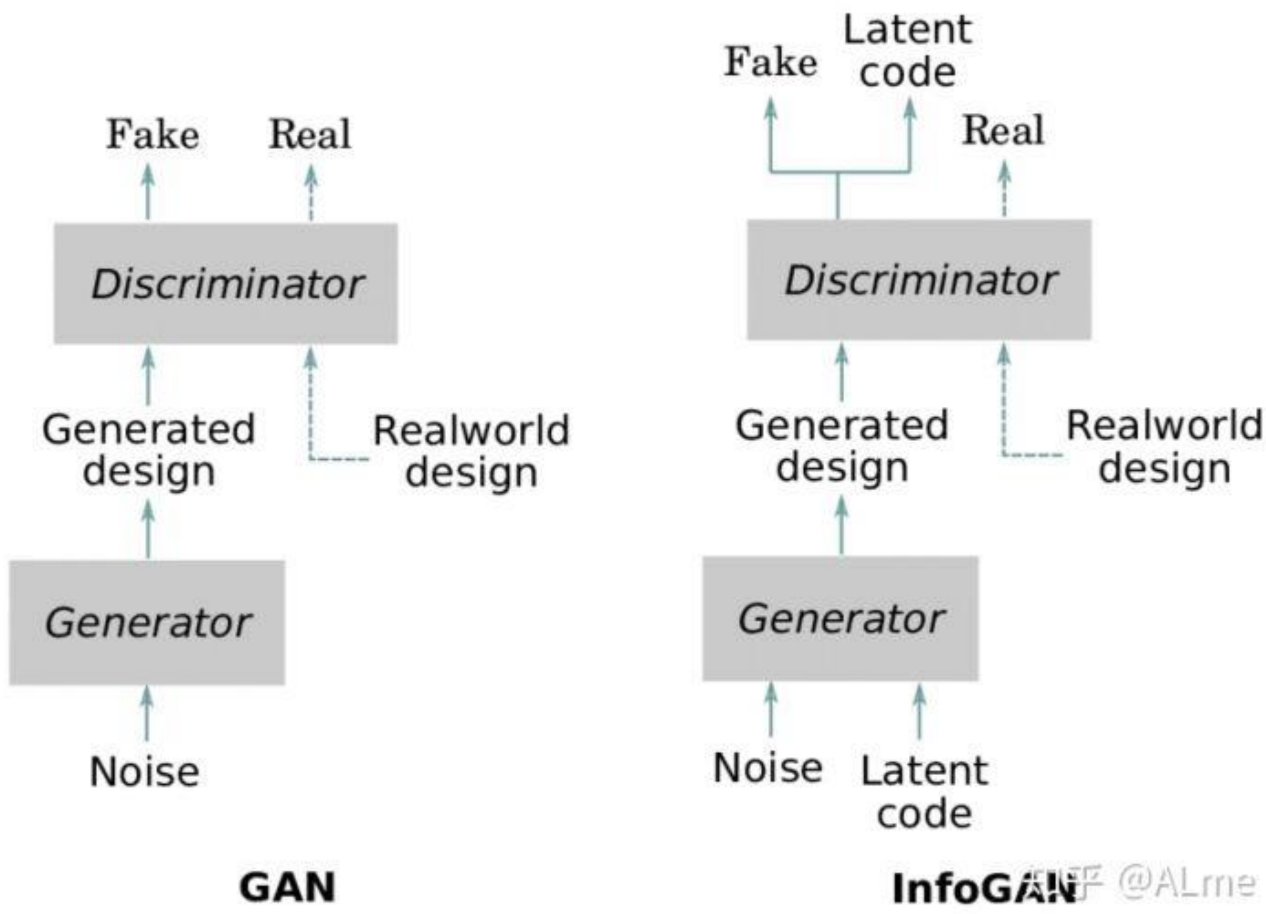
$$\begin{aligned}
I(c; G(z,c)) &= H(c) - H(c|G(z,c)) \\
&= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\
&= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\
&\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)
\end{aligned}$$

$$L_I(G,Q) = E_{c\sim P(c), x\sim G(z,c)}[\log Q(c|x)] + H(c)$$

Discriminator

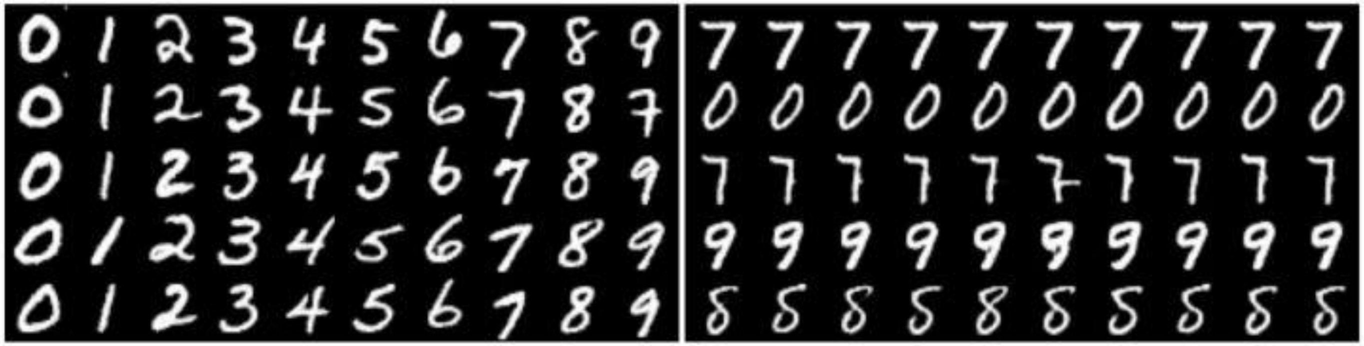
4.1. InfoGAN

- Architecture



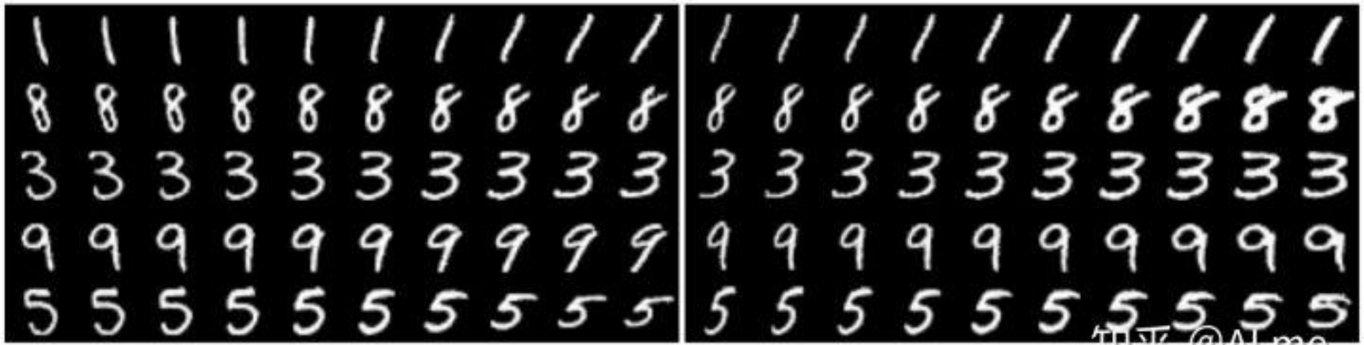
4.1. InfoGAN

- Disentangled Representation



(a) Varying c_1 on InfoGAN (Digit type)

(b) Varying c_1 on regular GAN (No clear meaning)



(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)

(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

4.2. BetaVAE

数据挖掘实验室
Data Mining Lab
LESS IS MORE

- VAE

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{KL}\big(q_\phi(\mathbf{z}|\mathbf{x}) \parallel p(\mathbf{z})\big)$$

- BetaVAE

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \underline{\beta}\, D_{KL}\big(\underline{q_\phi(\mathbf{z}|\mathbf{x})} \parallel p(\mathbf{z})\big)$$

Standard Gaussian distribution

4.2. BetaVAE

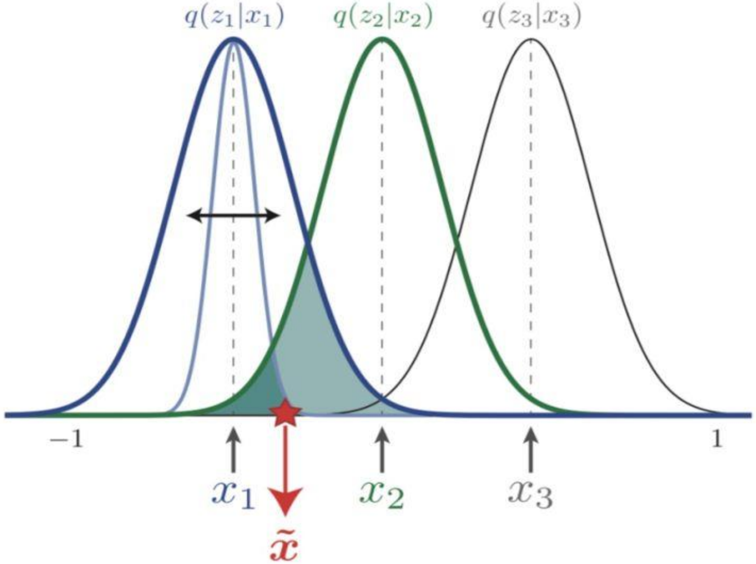
数据挖掘实验室
Data Mining Lab

- Information Bottleneck

$$\max[I(Z;\underline{Y}) - \beta I(\underline{X;Z})]$$

- Understanding BetaVAE

$$\mathcal{L}(\theta, \phi; \mathbf{x}, \mathbf{z}, \beta) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta\, D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}) \,\|\, p(\mathbf{z}))$$



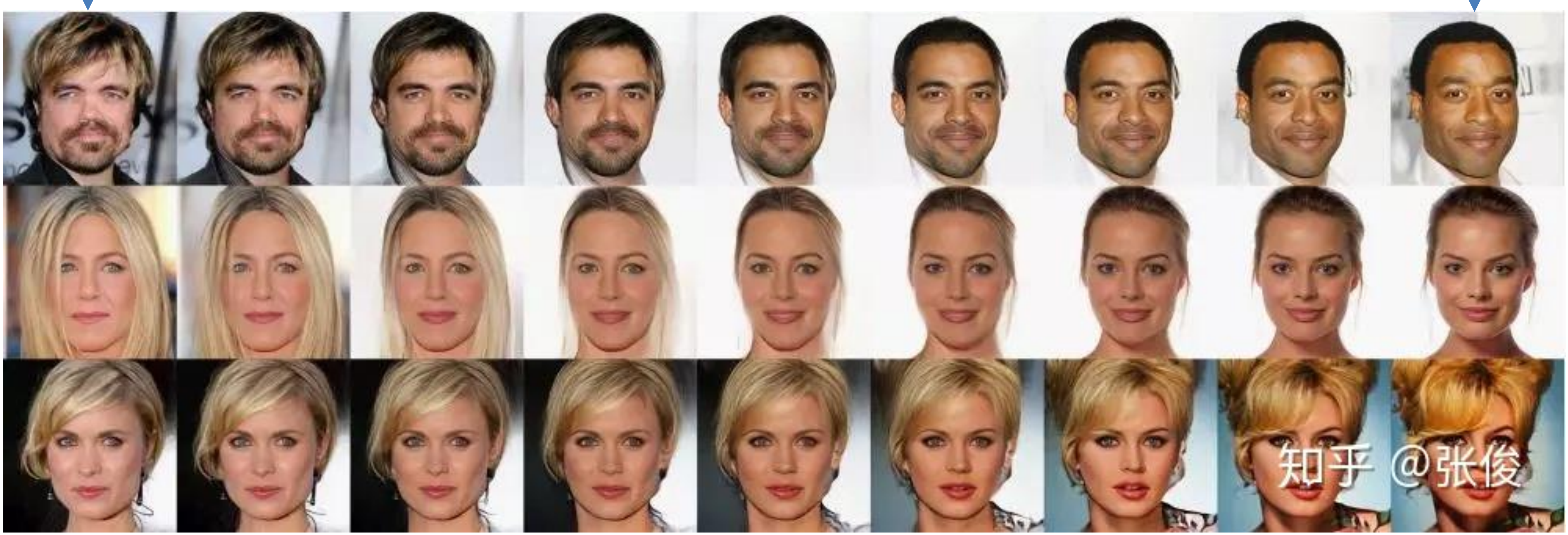
4.3. Standard Gaussian Distribution

- Decomposition

$$P(x, y) = p(x)p(y)$$

- Interpolation

$$z_1 + z_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$



Generalization Bound

5.1. Computational learning theory

数据挖掘实验室
Data Mining Lab
LESS IS MORE

- Generalization Error

$$E(h;\mathcal{D}) = P_{\boldsymbol{x}\sim\mathcal{D}}\big(h(\boldsymbol{x}) \neq y\big)$$

- Empirical error

$$\widehat{E}(h;D) = \frac{1}{m}\sum_{i=1}^{m}\mathbb{I}\big(h(\boldsymbol{x}_i) \neq y_i\big)$$

- Generalization bound

$$E(h) \leqslant \epsilon$$

5.1. Computational learning theory

数据挖掘实验室
Data Mining Lab
LESS IS MORE

- Probably Approximately Correct (PAC)

$$P(E(h) \leqslant \epsilon) \geqslant 1 - \delta \,, \quad 0 < \epsilon, \delta < 1$$

- Sample Complexity

$$m \geqslant \operatorname{poly}(1/\epsilon, 1/\delta, \operatorname{size}(\boldsymbol{x}), \operatorname{size}(c))$$

- Properly PAC learnable

$$\mathcal{H} = \mathcal{C}$$

- Infinity hypothesis space

$$|\mathcal{H}| = \infty$$

- If $|\mathcal{H}| \neq \infty$ and $c \in \mathcal{H}$

$$\begin{aligned} P\big(h(\boldsymbol{x}) = y\big) &= 1 - P\big(h(\boldsymbol{x}) \neq y\big) \\ &= 1 - E(h) \\ &< 1 - \epsilon \,. \end{aligned}$$

$$\begin{aligned} P\big((h(\boldsymbol{x}_1) = y_1) \wedge \ldots \wedge (h(\boldsymbol{x}_m) = y_m)\big) &= \big(1 - P\left(h\left(\boldsymbol{x}\right) \neq y\right)\big)^m \\ &< (1-\epsilon)^m \,. \end{aligned}$$

$$\begin{aligned} P\big(h \in \mathcal{H} : E(h) > \epsilon \wedge \widehat{E}(h) = 0\big) &< |\mathcal{H}|(1-\epsilon)^m \\ &< |\mathcal{H}| e^{-m\epsilon} \leqslant \delta \,, \end{aligned}$$

$$m \geqslant \frac{1}{\epsilon}\left(\ln|\mathcal{H}| + \ln\frac{1}{\delta}\right)$$

5.1. Computational learning theory

数据挖掘实验室 Data Mining Lab

- If $|\mathcal{H}| \neq \infty$ & $c \notin \mathcal{H}$

$$P\left(E(h) - \min_{h' \in \mathcal{H}} E(h') \leqslant \epsilon\right) \geqslant 1 - \delta$$

- If $|\mathcal{H}| = \infty$

$$\mathrm{VC}(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\}$$

$$P\left(E(h) - \widehat{E}(h) \leqslant \sqrt{\frac{8d\ln\frac{2em}{d} + 8\ln\frac{4}{\delta}}{m}}\right) \geqslant 1 - \delta\ .$$

5.2. Generalization bound via MI

- Low MI => Generalization

$$|\mathbf{E}\left[\phi_T - \mu_T\right]| \leq \sigma\sqrt{2I(T;\phi)},$$

$$|\mathbf{E}[\phi_T] - \mathbf{E}[\mu_T]| \leq \sigma\sqrt{\frac{2I(T;\phi)}{n}}.$$

$$\mathbf{E}[L(\hat{f}) - \hat{L}(\hat{f})] \leq \sqrt{\frac{I(\hat{f}(\mathbf{x});\mathbf{Y})}{2n}}.$$

5.3. Generalization bound via CMI

- $I(T(X), X) \to \infty$, but strong generalization holds

- CMI is finite

$$CMI_D(T) = I\left(T\left(\tilde{X}_S\right); S|\tilde{X}\right)$$

$$CMI_D(T) \leq H(S) = \log 2^n = n$$

- Low CMI => Generalization

- CMI is bounded by VC dimension and Compression Schemes

- CMI is bounded from (ε, δ)-DP (via TV Stability)

Thank you