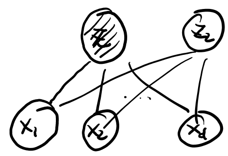
Lec 17: Restricted Boltzmann Machine (RBM)

July 8, 2020 12:51 AM

GMM: $p(x) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(x; \mu_k, S_k)$



MCMC: Markov Chain Monte Carlo

Gibbs sampling

Boltzmann Distribution:

r.v. $\underline{\vec{S} \in \{\pm 1\}^m}$ iff $\exists$ symmetric matrix $W \in \mathbb{S}^{(m+1)\times(m+1)}$

$$p(\vec{S} = \vec{s}) = \exp(\underline{\vec{s}^T W \vec{s}} - A(w))$$

$$= \underline{\sum_{i,j=1}^{m} s_i s_j W_{ij} + \sum_{i=1}^{m} s_i W_{i,m+1} + \sum_{j=1}^{m} s_j W_{m+1,j}}$$

$A(w)$: log-partition function: $\boxed{A(w) = \log \sum_{\vec{s} \in \{\pm 1\}^m} \exp(\vec{s}^T W \vec{s})}$ 2^m

BD $\in$ ExpFamily, $\underline{T(\vec{s}) = \vec{s}\,\vec{s}^T}$

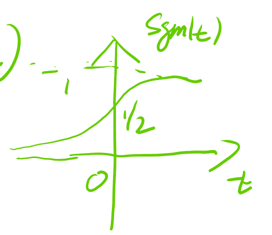
BD is $\underline{\text{intractable}}$

Example: $m=1$

$$\begin{cases} p(s=1) \propto \exp\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}^T \begin{pmatrix} 0 & W_{12} \\ W_{12} & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix}\right) \\ \qquad\quad = \exp(2W_{12}) \\ \Pr(s=-1) \propto \exp(-2W_{12}) \end{cases}$$

$$P(s=1) = \frac{\exp(2W_{12})}{\exp(2W_{12}) + \exp(-2W_{12})} = \frac{1}{1+\exp(-4W_{12})}$$

$$= \text{sgm}(4W_{12}), \qquad \text{sgm}(t) = \frac{1}{1+e^{-t}}$$



$$\bigcup_{W} BD_W \subsetneq \text{Prob}(\{\pm 1\}^m)$$

Boltzmann Machine (Ackley, Hinton & Sejnowski '85)

$\vec{S} = (\vec{x}, \vec{z})$, $\vec{x} \in \{\pm 1\}^d$, $\vec{z} \in \{\pm 1\}^t$ as $t \to 2^d$.

$$\underline{\underline{P_W((x, z))}} \propto \exp\left[\begin{pmatrix} x \\ z \\ 1 \end{pmatrix}^T \begin{pmatrix} W_{xx} & W_{xz} & b_x \\ W_{xz}^T & W_{zz} & b_z \\ \cdots & \cdots & \end{pmatrix} \begin{pmatrix} x \\ z \\ 1 \end{pmatrix}\right]$$

$$\underline{P_W(x,z)} \propto \exp\left[\begin{pmatrix} x \\ z \\ 1 \end{pmatrix}^T \underbrace{\begin{pmatrix} W_{xx} & W_{xz} & b_x \\ W_{xz}^T & W_{zz} & b_z \\ b_x^T & b_z^T & 0 \end{pmatrix}}_{W} \begin{pmatrix} x \\ z \\ 1 \end{pmatrix} \right]$$

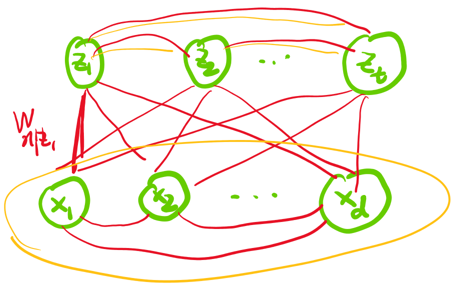
Given $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n$, learn $\underline{P_W(\vec{x})}$.

Restricted Boltzmann Machine (Smolensky '86)

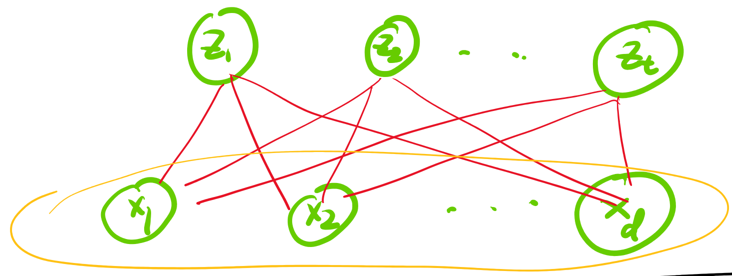
$$P_W(x,z) \propto \exp\left[\begin{pmatrix} x \\ z \\ 1 \end{pmatrix}^T \begin{pmatrix} \cancel{W_{xx}}\ 0 & W_{xz} & b_x \\ W_{xz}^T & \cancel{W_{zz}}\ 0 & b_z \\ b_x^T & b_z^T & 0 \end{pmatrix} \begin{pmatrix} x \\ z \\ 1 \end{pmatrix} \right]$$

$$= \exp\left(x^T W_{xz} z \right)$$

BM



RBM



Given $\vec{x}_1, \vec{x}_2, \dots, \vec{x}_n \sim P_W(\vec{x})$, estimate W in $P_W(x,z)$

① EM $\longrightarrow$ $A(W)$

② ML $\quad \min_W KL(\hat{p} \,\|\, P_W(x))$

$$\equiv \min_W -\frac{1}{n}\sum_{i=1}^n \log P_W(\vec{x}_i)$$

$s_i = (x_i, z)$

$$\equiv \min_W -\frac{1}{n}\sum_{i=1}^n \log \sum_{z\in\{\pm1\}^t} \underbrace{e^{\vec{s}_i^T W s_i - A(W)}}_{P_W(x_i,z)}$$

$$\frac{\partial}{\partial W} = -\frac{1}{n}\sum_{i=1}^n \frac{\sum_{z\in\{\pm1\}^t} P_W(x_i,z)\left[s_i s_i^T - \nabla A(W)\right]}{P_W(x_i)}$$

$P_W(x_i)$

$$= -\frac{1}{n}\sum_{i=1}^n \sum_{z\in\{\pm1\}^t} P_W(z|x_i)\left[\vec{s}_i\vec{s}_i^T - \nabla A(W)\right]$$

$$\nabla A(W) = \mathbb{E}_{P_W(x,z)} T(s) = \mathbb{E}_{P_W(x,z)} \vec{s}\vec{s}^T$$

$$\sum_{z \in \{\pm 1\}^t} \qquad \sum_{P_w(x,z)} SS^T$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \sum_{z\in\{\pm1\}^t} P_w(z|x_i)\cdot \vec{s}_i\vec{s}_i^T + \nabla A(w)$$

$$= -E_{\hat{P}_w} \vec{s}\vec{s}^T + E_{P_w} \vec{s}\vec{s}^T$$

$$\hat{P}_w(X = x_i, Z = z) := \frac{1}{n}\cdot P_w(z|x_i)$$

$$E_{P_w} \vec{s}\vec{s}^T \approx \frac{1}{n}\sum_{i=1}^{n} \vec{s}_i\vec{s}_i^T, \quad \vec{s}_i \sim P_w(s)$$

Gibbs sampling: $\vec{S} \sim P_w(s)$

initialize $\vec{s} \in \{\pm 1\}^m$

repeat {
for $j = 1, 2, \cdots, m$

$p_j = P_w(S_j = 1 \mid \vec{s}_{\setminus j})$

set $s_j = 1$ w.p. p_j and $s_j = -1$ otherwise

$$P_w(\vec{S} = \vec{s}) \propto \exp(\vec{s}^T W \vec{s})$$

$$P_w(s_j = 1 \mid \vec{s}_{\setminus j}) = \frac{P_w(s_j = 1, \vec{s}_{\setminus j})}{P_w(\vec{s}_{\setminus j})} = \frac{P_w(s_j=1, \vec{s}_{\setminus j})}{P_w(s_j=1, \vec{s}_{\setminus j}) + P_w(s_j=-1, \vec{s}_{\setminus j})}$$

$$= \mathrm{sgm}(4\langle w_{\setminus j, j}, \vec{s}_{\setminus j}\rangle)$$

$$P_w(z_j = 1 \mid \vec{z}_{\setminus j}, \vec{x}) = \mathrm{sgm}(2\langle w_{:j}, \vec{x}\rangle)$$

$$P_w(x_j = 1 \mid \vec{x}_{\setminus j}, \vec{z}) = \mathrm{sgm}(2\langle w_{j,:}, \vec{z}\rangle)$$

For $t = 1, 2, \cdots$

For $t = 1, 2, \cdots$

Sample $x_i \sim$ training data set

sample $(\vec{X}, \vec{Z}) \sim P_W(x, z)$

sample $\vec{z}_i \sim P_W(z \mid x_i)$

$$\widehat{\frac{\partial}{\partial W}} = -\begin{pmatrix} x_i \\ z_i \end{pmatrix}\begin{pmatrix} x_i \\ z_i \end{pmatrix}^T + \begin{pmatrix} \vec{X} \\ \vec{Z} \end{pmatrix}\begin{pmatrix} \vec{X} \\ \vec{Z} \end{pmatrix}^T$$

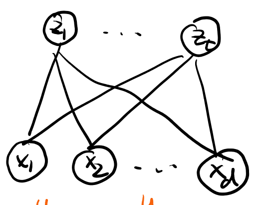
$$P_W(x, z) \propto \exp\left(\begin{pmatrix} x \\ z \\ 1 \end{pmatrix}^T W \begin{pmatrix} x \\ z \\ 1 \end{pmatrix} \right)$$

draw $\begin{pmatrix} x \\ z \end{pmatrix} \sim P_W$

For RBM: $P_W(x) \propto \exp(x^T W_{:,t+1}) \cdot \prod_{j=1}^{t} \left[e^{x^T W_{:j}} + e^{-x^T W_{:j}} \right]$

RBM is a stochastic NN

deterministic NN



$$y_j = \Pr(Z_j = 1 \mid \vec{X} = \vec{x}) = \text{sgm}(2 W_{:j}^T \vec{X})$$

$$\begin{cases} \vec{h} = 2 W^T \vec{X} \\ \vec{y} = \text{sgm}(\vec{h}) \end{cases}$$