Lec 18: Belief Networks (BN)

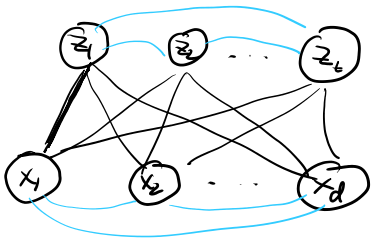
July 8, 2020 10:03 AM

Recap: BM / RBM

parameterization: $P_W(s_1, s_2, \cdots, s_m) \propto \exp\left(\vec{s}^T W \vec{s} - A(W)\right)$

Gibbs sampling: $P_W(s_i \mid s_{\setminus i}) = \mathrm{sgm}\left(4 s_i \langle W_{ij}, s_j \rangle\right)$

Graph: RBM

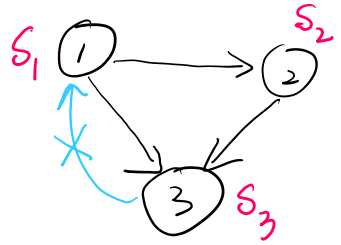


undirected graph

W is symmetric

DAG: directed acyclic graph

$pa(v) = \{u : \vec{uv} \in E\}$



$pa(1) = \phi$

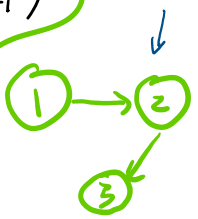
$pa(2) = 1$

$pa(3) = \{1, 2\}$

Belief Net (BN) / Bayes: $\vec{S} = (s_1, s_2, \cdots, s_m)$

$$p(s_1, s_2, \cdots, s_m) = \prod_{j=1}^{m} p(s_j \mid s_1, s_2, \cdots, s_{j-1}) \approx \prod_{j=1}^{m} p(s_j \mid pa(s_j))$$

$p(s_1, s_2, s_3) = p(s_1)\, p(s_2 \mid s_1)\, p(s_3 \mid s_1, s_2)$



$p(s_1, s_2, s_3) = p(s_1)\, p(s_2 \mid s_1)\, p(s_3 \mid s_2)$

Thm: Fix a DAG. Given $\{p(s_j \mid pa(s_j))\}_{j=1}^m$, can construct

$$p(s_1, s_2, \cdots, s_m) = \prod_{j=1}^{m} p(s_j \mid pa(s_j)).$$

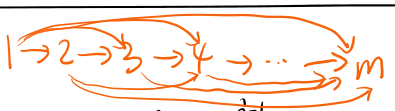
$\times \begin{cases} p(s_1) \\ p(s_2 \mid s_1) \\ p(s_3 \mid s_2) \end{cases}$

Ex: If the graph is not a DAG, then thm is not true.

Rem: $s \in \{\pm 1\}^m$, need $2^m - 1$ parameters to specify $p(s_1, s_2, \cdots, s_m)$

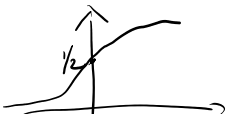
DAG with in-degree $\le$ max K: $(2^K - 1)\, m$

Sigmoid BN (Neal '92)

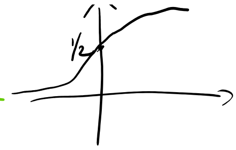


specify $P_W(s_j \mid s_{<j}) = \mathrm{sgm}\left(s_j \cdot \left(\sum_{k=1}^{j-1} W_{jk} s_k + b\right)\right)$

$\mathrm{Sgm}(t) = \frac{1}{1 + e^{-t}}$



specify $P_W(s_j \mid s_{<j}) = \mathrm{sgm}\left(s_j \cdot \left(\sum_{k=1}^{j-1} W_{jk} s_k + b\right)\right)$



$= \mathrm{sgm}(s_j \cdot W_{j:} \vec{s})$

$W \in \mathbb{R}^{m \times (m+1)}$ is lower-triangular

$W_{jk} = 0$ if $j \le k \le m$

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & b_1 \\ \times & 0 & 0 & 0 & 0 & b_2 \\ \times & \times & 0 & 0 & 0 & \vdots \\ \times & \times & \times & 0 & 0 & \\ \times & \times & \times & \times & 0 & b_m \end{bmatrix}$$

$$P_W(s_1, s_2, \cdots, s_m) = \prod_{j=1}^{m} P_W(s_j \mid s_{<j}) = \prod_{j=1}^{m} \mathrm{sgm}(s_j \cdot W_{j:} \vec{s})$$

Universality: $S = (X, Z)$, $P_W(x)$ can approx any discrete distr. for BN.

SBN: Max Likelihood

Given $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_n \in P(X)$, $\min_W KL(\hat{P}(x) \| P_W(x))$

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

$$\frac{\partial}{\partial W_{jk}} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial_{W_{jk}} P_W(\vec{x}_i)}{P_W(\vec{x}_i)} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial_{W_{jk}} \sum_{\vec{z}_i} P_W(\vec{x}_i, \vec{z}_i)}{P_W(\vec{x}_i)}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \sum_{\vec{z}_i} \frac{\partial_{W_{jk}} P_W(\vec{x}_i, \vec{z}_i)}{P_W(\vec{x}_i)}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \sum_{\vec{z}_i} \frac{P_W(\vec{x}_i, \vec{z}_i)}{P_W(\vec{x}_i)} \cdot \partial_{W_{jk}} \log P_W(\vec{x}_i, \vec{z}_i)$$

$\frac{P_W(\vec{x}_i, \vec{z}_i)}{P_W(\vec{x}_i)} = P_W(\vec{z}_i \mid \vec{x}_i)$, $\quad P_W(\vec{x}_i, \vec{z}_i) = \prod_j \mathrm{sgm}(s_j \cdot W_{j:} \vec{s})$

$$= -E_{\hat{P}_W(s)} \partial_{W_{jk}} \log \prod_j \mathrm{sgm}(s_j \cdot W_{j:} \vec{s}), \quad \hat{P}_W(\vec{s}) = \hat{P}_W(\vec{x}, \vec{z}) = \hat{p}(\vec{x}) \cdot P_W(\vec{x} \mid \vec{z})$$

$$= -E_{\hat{P}_W} \partial_{W_{jk}} \log \mathrm{sgm}(s_j W_{j:} \vec{s})$$

$W_{j:}\vec{s} = s_j \sum_k W_{jk} s_k$

$$= -E_{\hat{P}_W} \frac{\mathrm{sgm}'(s_j W_{j:} \vec{s})}{\mathrm{sgm}(s_j W_{j:} \vec{s})} \cdot s_j \cdot s_k$$

$\mathrm{sgm}'(t) = \mathrm{sgm}(t) \cdot \mathrm{sgm}(-t)$

$$= -E_{\hat{P}_W} \operatorname{sgn}(-s_j W_{j:} \vec{s})\, s_j s_k$$

Gibbs sampling: sample x from training data

$$P(z_j = z \mid z_{\setminus j}, x) = \operatorname{sgm}(z W_{z:} \vec{s}) \cdot \prod_{k>j} \operatorname{sgn}\left(s_k [W_{k:} s + W_{kz}(z - s_j)]\right)$$

$$\propto P(z_j = z, z_{\setminus j}, x) = \prod_k P(s_k \mid \text{pa}(s_k))$$

Ex: Can we use EM?

DBN (Bengio & Bengio '99)

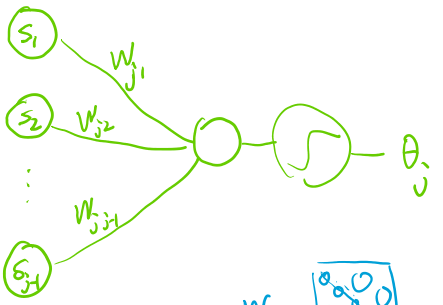
— specify m univariate densities $P(s_j \mid s_{<j}) = P_j(s_j ; \theta_j(s_{<j}))$ Gaussian

for SBN: Bernoulli for each $P_j(s_j = 1) = \theta_j(s_{<j})$

— $\theta_j(s_{<j}) = \operatorname{sgm}\left(\sum_{k<j} W_{jk} s_k\right)$

— DBN

$$\begin{cases} \vec{h}_0 = \vec{s} \\ \vec{h}_\ell = \sigma(W_\ell \vec{h}_{\ell-1}), \quad \ell = 1, 2, \cdots, L-1 \\ \vec{\theta} = \sigma(W_L \vec{h}_{L-1}) \end{cases}$$



θ_j can only depend on $s_1, s_2, \cdots, s_{j-1}$