Lec 19: Learning Graphical Models

- Parameter learning:
  Graph fixed, learning params of BN, MN

- Structure learning:
  find the graph ie how variables depend on each other

In PAC learning theory, examples labelled by unknown function.
Goal was to to find that function

We have some samples, we want to learn the joint distribution

Maximum likelihood

dataset D sampled (iid) from p^*
Goal: find dist p which is "close" to p^*
closeness of pr. distributions:
KL divergence

$$D_{KL}(p^* \| p) = \sum_x p^*(x) \log\left(p^*(x)/p(x)\right)$$

$$= \sum_x p^*(x) \log p^*(x) - \sum_x p^*(x) \log p(x)$$

$$= -H(p^*) - \mathbb{E}_{x\sim p^*} \log p(x)$$

entropy of p^*

- p^* is unknown, we cannot compute $H(p^*)$
- Same as maximize $\mathbb{E}_{x\sim p^*} \log p(x)$

- $\mathbb{E}_{x\sim p^*} \log p(x)$ is "log likelihood"
  find dist p that maximizes the probability of the dataset

- Compute by taking average over samples

$$\mathbb{E}_{x\sim p^*} \log p(x) \approx \frac{1}{|D|} \sum_{x\in D} \log p(x)$$

D is iids from p^*

Max. Lik. learning

$$\max_{p\in M} \frac{1}{|D|} \sum_{x\in D} \log p(x)$$

Empirical Distribution

$$\tilde{p}(x) = \frac{1}{|D|} \sum_{y\in D} \delta(x,y)$$

Let w be parameters of the dist. to be learnt.

$\{x_1 \cdots x_N\} = D$

Theorem:

$$\arg\max_w \frac{1}{N} \sum_{i=1}^{N} \log p(x_i \mid w)$$

$$= \arg\min_w D_{KL}(\tilde{p}(x) \| p(x|w))$$

Proof:

$$\text{RHS} = \sum_x \tilde{p}(x) \log \tilde{p}(x) - \sum_x \tilde{p}(x) \log p(x|w)$$

$-H(\tilde{p})$ does not depend on w

$$= \max_w \sum_x \tilde{p}(x) \log p(x|w)$$

$$= \max_w \sum_x \frac{1}{|D|} \sum_{y\in D} \delta(x,y) \log p(x|w)$$

$$= \max_w \frac{1}{|D|} \sum_{y\in D} \log p(y|w)$$

General loss function

$$\mathbb{E}_{x\sim p^*} [\mathcal{L}(x,p)]$$

$$\mathcal{L}(x,p) = -\log p(x)$$

Conditional Random Fields (CRFs)

x, y

Interested in finding $p(y|x)$

$$\mathcal{L}(x,y,p) = -\log p(y|x)$$

Max. Lik. Learning for BN

$$p_\theta(x) = \prod_{i=1}^{n} \theta_{x_i \mid par(x_i)}$$

$$D = \{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$$

$$\mathcal{L}(\theta, D) = -\sum_{i=1}^{m} \log p_\theta(x^{(i)})$$

$$= -\sum_{i=1}^{m} \sum_{j=1}^{n} \log \theta_{x_j^{(i)} \mid x^{(i)}_{par(j)}}$$

$$= -\sum_{j=1}^{n} \left(\sum_{x_{par(j)}} \sum_{x_j} \#(x_j, x_{par(j)}) \log \theta_{x_j^{(i)} \mid x^{(i)}_{par(j)}} \right)$$

minimized when for each $j \in \{1, \ldots n\}$

$$\sum_{x_{par(j)}} \sum_{x_j} \#(x_j, x_{par(j)}) \log \theta_{x_j \mid x_{par(j)}}$$

minimized at

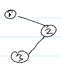
$$\theta^*_{x_j \mid par(x_j)} = \frac{\#(x_j, x_{pa(j)})}{\# x_{pa(j)}}$$

empirical value for $p_{x_j \mid par(x_j)}$

Learning in Markov Networks

$$p_\theta(x_1, \ldots x_n) = \frac{1}{Z(\theta)} \prod_{c\in C} \phi_c(x_c, \theta)$$

$$p(x_1,x_2,x_3) = \frac{1}{Z} \phi_{12}(x_1, x_2)\, \phi_{23}(x_2\, x_3)$$



$x_1, x_2, x_3 \in \{0,1\}$

$4 + 4 = 8$

Suppose there are m edges,
$\Rightarrow 4m$ parameters

$\phi_{12}(00) = \theta_{12,00}$

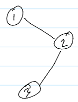
$\phi_{12}(01) = \theta_{12,01}$

$\vdots$

$\phi_{23}(00) = \theta_{23,00}$

$\vdots$

$\phi_{23}(11) = \theta_{23,11}$



$$\frac{1}{Z(\theta)} \exp\left(\sum_{c\in C} \log \phi_c(x_c; \theta)\right)$$

$$= \frac{1}{Z(\theta)} \exp(\underline{\tilde{\theta}^T \cdot f(x)})$$

$4m$ dim vector

$x = (0\ 1\ 0)$ $\quad y$

$$f(010) = \begin{array}{c} \phi_{12} \\ \\ \\ \\ \phi_{23} \end{array} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} \begin{array}{c} 00 \\ 01 \\ 10 \\ 11 \\ 00 \\ 01 \\ 10 \\ 00 \end{array} \qquad \begin{bmatrix} \log \theta_{12,00} \\ \vdots \\ \log \theta_{12,11} \\ \log \theta_{23,00} \\ \vdots \\ \log \theta_{23,11} \end{bmatrix} = \tilde{\theta}$$

x, x'

$f(x) = f(x')$

$$p(010) = \frac{1}{Z(\theta)} e^{(\tilde{\theta}_2 + \tilde{\theta}_7)}$$

$f(x)$: sufficient statistics

"exponential families"

Log Likelihood (LL) $= \log p(D, \theta)$

$$= \frac{1}{|D|} \sum_{x\in D} \log p(x : \theta)$$

$$= \left(\frac{1}{|D|} \sum_{x\in D} \theta^T \cdot f(x)\right) - \log(Z(\theta))$$

linear in θ

Learning by gradient descent

$$\nabla_\theta (LL) = \frac{1}{|D|} \sum_{x\in D} f(x) - \nabla_\theta \log Z(\theta)$$

sample mean of sufficient statistics

θ^*

$$\log Z(\theta) = \log \sum_x \exp(\theta^T \cdot f(x)) \qquad (e^{f(x)})' = e^{f(x)} \cdot f(x)$$

$$\nabla_\theta \log Z(\theta) = \frac{\sum_x \exp(\theta^T \cdot f(x))\, f(x)}{Z(\theta)}$$

$$= \sum_x p_\theta(x) f(x) = \mathbb{E}_{x\sim p_\theta(x)} f(x)$$

expected value of $f(x)$ according to p_θ

When $\nabla_\theta (LL) = 0$

sample mean of sufficient statistics = mean according MN of suff. statistics

"Moment matching of sufficient statistics."

$$\mathbb{E}_{x\sim p_\theta(x)} f(x) \approx \frac{1}{|S|} \sum_{y\in S} f(y)$$

Can generate samples from $p_\theta(x)$ by MCMC method