Directed Graphical Models

Monday, May 6, 2019 2:08 PM

Suppose we have 4 R.V.s (discrete) X, Y, W, Z we observe Z

we want $\mathbb{P}\{X = x \mid Z = z\}$

<u>Inference</u> calculating conditional, marginal, or joint dist^{n}s

<u>Learning</u> estimating parameters and dependence structure

$$P_{X|Z}(x|z) = \sum_{w,y} P_{X,W,Y|Z}(x,w,y|z) \quad O(|\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{W}|)$$

$x \in \mathcal{X}$ $\quad y \in \mathcal{Y}$ $\quad w \in \mathcal{W}$

$$P_{X,W,Y,Z} = P_{X|W,Y,Z} \cdot P_{W|Y,Z} \cdot P_{Y|Z} \cdot P_Z$$

Suppose that we knew $X, Y \perp\!\!\!\perp Z \mid W$ (X, Y are conditionally independent of Z given W) and $X \perp\!\!\!\perp W \mid Y$

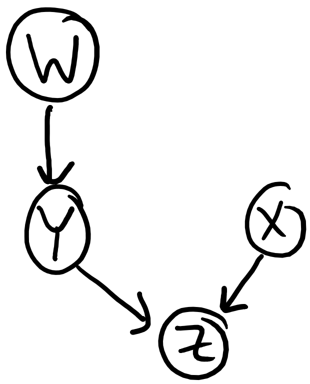
$$P_{X,W,Y|Z} \overset{?}{=} P_{X|Y} \cdot P_{Y|W} \cdot P_{W|Z}$$

$$= P_{X|W,Y,Z} \cdot P_{Y|W,Z} \cdot P_{W|Z}$$

$$\sum_{w,y} P_{X,W,Y|Z} = \sum_{w,y} P_{X|Y} \cdot P_{Y|W} \cdot P_{W|Z}$$

$$= \underbrace{\sum_{y} P_{X|Y} \Big(\underbrace{\sum_{w} P_{Y|W} P_{W|Z}}_{O(|\mathcal{Y}| \times |\mathcal{W}|)} \Big)}$$

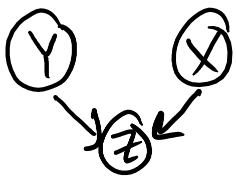
$$O(|\mathcal{X}|\cdot|\mathcal{Y}|)$$



__def__ A Bayesian network is a directed acyclic graph (DAG) s.t. vertices are variables and each vertex is conditionally independent from its non-descendents given its parents.

$Z \perp\!\!\!\perp W \mid X, Y$ __not__ necessarily the case that $\underline{Y} \perp\!\!\!\perp \underline{Z} \mid W$

descendent

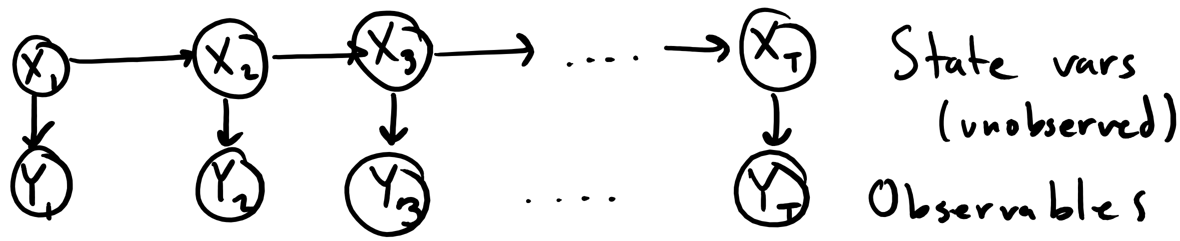


$\Rightarrow Y \perp\!\!\!\perp X$

$$\sum_z P_{Z|X,Y} \cdot P_X \cdot P_Y = P_X \cdot P_Y$$
$$= P_{X,Y}$$

Hidden Markov Models

Monday, May 6, 2019 2:09 PM



$$P_{X_{1:T}, Y_{1:T}} = P_{X_1} \cdot P_{Y_1|X_1} \cdot P_{X_2|X_1} \cdot P_{Y_2|X_2} \cdot P_{X_3|X_2} \cdot P_{Y_3|X_3} \cdot \ldots \cdot P_{X_T|X_{T-1}} \cdot P_{Y_T|X_T}$$

$$= P_{X_1} \cdot P_{Y_1|X_1} \cdot \prod_{t=2}^{T} P_{X_t|X_{t-1}} \cdot P_{Y_t|X_t}$$

Maximum likelihood estimation $\left(P_{X_t|X_{t-1}, \theta}, \; P_{Y_t|X_t, \theta} \right)$

$$\max_\theta \; \log \sum_x P_{X,Y|\theta}(x,y|\theta) = \log P_{Y|\theta}(y|\theta)$$

all time

$$= \log \sum_x q(x) \cdot \frac{P_{Y,X|\theta}(y,x|\theta)}{q(x)}$$

some other distn $\quad q(x) > 0$

$$\geq \sum_x q(x) \log \frac{P_{Y,X|\theta}(y,x|\theta)}{q(x)} \qquad \text{(Jensen's inequality)}$$

$$=: F(q, \theta)$$

$$\operatorname*{argmax}_q F(q, \theta) = P_{X|Y,\theta}(x|y;\theta)$$

E-step is $q_{k+1}(x) \leftarrow P_{X|Y,\theta_k}(x|y,\theta_k)$

M-step is $\max_{\theta} \sum_x \underbrace{P_{X|Y,\theta_k}(x|y,\theta_k)}_{q_{k+1}(x)} \log P_{X,Y|\theta}(x,y|\theta)$

<u>HMM (discrete)</u> Φ_{ij} is $\mathbb{P}$ of transition from $X_t = j$ to $X_{t+1} = i$

$$X_t = (0,\ldots,0,\overset{i}{1},0,\ldots,0) \text{ - use one-hot encoding}$$

$$P_{X_t|X_{t-1}}(x_t|x_{t-1}) = \prod_i \prod_j \Phi_{ij}^{x_{t,i}\, x_{t-1,j}}$$

$$\log P_{X_t|X_{t-1}}(x_t|x_{t-1}) = \sum_{i,j} x_{t,i}\, x_{t-1,j} \log \Phi_{ij}$$

$$= x_t^T (\log \Phi)\, x_{t-1}$$

Similarly, $\log P_{Y_t|X_t}(y_t|x_t) = y_t^T (\log \overset{\substack{\text{emission matrix}\\ \downarrow}}{E})\, x_t$

$$\log p_{X,Y|\theta} = \log p_{X_1} + \sum_{t=1}^{T} \log P_{Y_t|X_t} + \sum_{t=2}^{T} \log P_{X_t|X_{t-1}}$$

M-step: $\langle g(x) \rangle_{\theta_k} = \sum_x P_{X|Y,\theta_k}\, g(x)$

calculate at max $\langle \log P_{X,Y|\theta} \rangle_{\theta_k}$

$$\max_{\Phi} \left\langle \sum_{t=2}^{T} x_t^T (\log \Phi) x_t \right\rangle_{\theta_k} \text{ s.t. } \sum_i \Phi_{ij} = 1$$

solution is $\Phi_{ij} \leftarrow \dfrac{\sum_{t=2}^{T} \langle X_{t,i} X_{t-1,j} \rangle_{\theta_k}}{\sum_{t=2}^{T} \langle X_{t-1,j} \rangle_{\theta_k}}$

Similar calculations for other parameters.

ex $\langle X_{t,i} \rangle_{\theta_k} = \dfrac{P_{X_t, Y | \theta_k}(i, y)}{\sum_j P_{X_t, Y | \theta_k}(j, y)}$

$P_{X_t, Y}(i, y) = \underbrace{P_{X_t, Y_{1:t}}(i, y_{1:t})}_{\alpha_{t,i}} \cdot \underbrace{P_{Y_{t+1:T} | X_t}(y_{t+1:T} | i)}_{\beta_{t,i}}$

(E-step)

$\alpha_{t,i} = P_{X_t, Y_{1:t}}(i, y_{1:t}) = \left[\sum_j P_{X_{t-1}, Y_{1:t-1}}(j, y_{1:t-1}) \cdot P_{X_t | X_{t-1}}(i | j) \cdot\right] \cdot P_{Y_t | X_t}(y_t | i)$

$= \left[\sum_j \alpha_{t-1,j} \; P_{X_t | X_{t-1}}(i | j)\right] \cdot P_{Y_t | X_t}(y_t | i)$

↳ forward pass for $t = 1, 2, \dots, T$

Similar for β but backward pass.

<u>Other tasks</u>

<u>decoding</u> find the most likely state sequence for known parameters and observed Y's

$\underset{x_{1:T}}{\operatorname{argmax}} \; P_{X|Y}(x_{1:T} | y_{1:T}) = \underset{x_{1:T}}{\operatorname{argmax}} \; P_{X,Y}(x, y)$

Maximum A-Posteriori (MAP)

$$= \underset{k}{\operatorname{argmax}} \max_{x_{1:T-1}} p(x_T = k, x_{1:T-1}, y_{1:T})$$

Viterbi algorithm recursively computes

EM for HMM (discrete) is Baum-Welsh