MRFs and HMMs

MRF refers to "Markov Random Field".

HMM refers to "Hidden Markov Model".

Def: i) $(X_s)_{s\in G}$ collection of r.v's is a random field indexed by G. set of nodes of graph.

ii) $s \sim t$ for $s, t \in G$ means they're neighbour.

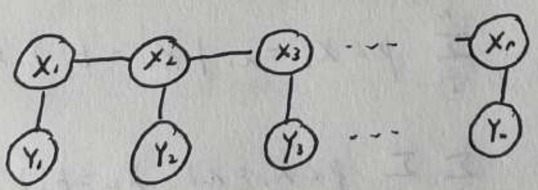
$N(t) = \{ s \in G \mid s \sim t \}$.

iii) $(X_s)_{s\in G}$ is MRF if $p(X_t = x_t \mid X_s = x_s,\ s \neq t)$

$= p(X_t = x_t \mid X_s = x_s,\ s \in N(t)) \stackrel{\Delta}{=}: p(x_t \mid x_{N(t)})\ \forall t.$

iv) HMM is a MRF, s.t. some r.v.'s are observed but others are hidden.

e.g.



Markov Chain (Hidden)

Observable process.

$$p(y_t \mid x, y_{\neq t}) = p(y_t \mid x_t).$$

HMMs fit in Bayesian Framework nicely:

For X unknown: prior: $Y \mid X \xrightarrow[\text{Formula}]{\text{Bayesian}}$ Posterior: $X \mid Y$.

we will make a reseasonable estimate for $p(X \mid Y)$ from $Y \mid X$.

(1) Gibbs Dist.:

Next. We will specify a MRF. by 2 methods:

i) Set of Conditional Dist:

$(p(x_t | x_{N(t)}))_{t\in G}$ may not be a consistent prob. dist. when we give MRF one dist.

Thm. If we have specified condition dist. $\mathcal{L}(X_1|X_2)$ for r.v. X_1, X_2. $X_1 \in S_1 = (a_i)_1^n$. $X_2 \in S_2 = (b_i)_1^m$. Then. We're free to decide one more dist. $\mathcal{L}(X_2|X_1=a)$. $a\in S$.

Pf: By Bayesian Formula:

$$p(X_1=a_i | X_2=b_j) = p(X_1=a_i, X_2=b_j) / \sum_k^n p(X_1=a_k, X_2=b_j)$$

for $1\le i\le n$, $1\le j\le m$

With: $\sum_i^n p(X_i=a_i | X_2=b_j) = 1$. $\forall\ 1\le j\le m$.

$$\sum_i \sum_j p(X_1=a_i, X_2=b_j) = 1.$$

Consider $(p(X_1=a_i, X_2=b_j))_{i,j}$ as set of unknown variable. $p(a_i|b_j)$ is known. The order of liner equation above is $mn-m+1$. $\Rightarrow$ At most choose $\mathcal{L}(X_2|X_1=a)$

ii) Hammersley - Clifford Thm:

Def: i) Set of nodes $\mathcal{G}$ is complete if every distinct nodes are neighbor of each other.

ii) A clique is max set of nodes. s.t. complete.

iii) $\mathcal{G}$ is finite graph. Gibbs dist. w.r.t $\mathcal{G}$ is pmf $p(x) = \prod_{c \text{ is complete}} V_c(x)$. V_c only depends on $x_c = (x_s)_{s \in c}$. for c is clique. $x \in S_{\mathcal{G}}$. (config)

Rmk: i) $x_c = \eta_c \Rightarrow V_c(x) = V_c(\eta)$.

ii) $p(x)$ can be reduced : $\prod_{c \text{ clique}} V_c(x)$.

Thm. (H-C Thm)

$X = (X_1, X_2 \cdots X_n)$ has positive joint pmf. Then:

X is MRF on $\mathcal{G} \Leftrightarrow X$ has a Gibbs dist. on $\mathcal{G}$.

Pf: $S_{\mathcal{G}} = S_1 \times S_2 \cdots \times S_n$. S_k is state space of X_k.

Denote: 0 means arbitrary element. (fix)

($\Leftarrow$). Show: $p(x_t \mid x_{\neq t}) / p(0_t \mid x_{\neq t})$ only depends on $x_{N(t)}$.

Note: $p(x_t, x_{\neq t}) / p(0_t, x_{\neq t}) = p(x_t|\square)/p(0_t|\square)$

$$= \frac{\prod_{t \in c} V_c(x_t, x_{\neq t})}{\prod_{t \in c} V_c(0_t, x_{\neq t})} \cdot \frac{\prod_{t \notin c} V_c(x_t, x_{\neq t})}{\prod_{t \notin c} V_c(0_t, x_{\neq t})}$$

$$= \prod_{t \in c} V_c(x_t, x_{\neq t}) / \prod_{t \in c} V_c(0_t, x_{\neq t}).$$

($\Rightarrow$) We want to write $p(x)$ in form:

$\prod_A V_A(x)$. $V_A \equiv 1$. if A isn't complete.

Set: $p(x_D, 0_{D^c}) = \prod_{A \subseteq D} V_A(x)$. $D \subseteq \{1, 2, \cdots n\}$.

Then we can find V_c reccurrsively.

1) $D=\emptyset$. $p(0) = V_\emptyset(x)$

2) $D=\{t\}$. $V_{\{t\}}(x) = p(x_t, 0_{\neq t})/p(0)$

3) $D \subseteq \{1,2,\cdots n\}$. $V_D(x) = p(x_D, 0_{\neq D}) / \prod_{A \subsetneq D} V_A(x)$.

Next. prove: $V_A \equiv 1$. if A not complete.

By induction on $|A|$. $|A|=1$ ✓.

For $n=k+1$. (suppose $n \le k$ holds)

if $t, u \in A$. not neighbour. $A = \{t,u\} \cup B$

Note: $p(x_A, 0_{\neq A}) = p(x_t, x_u, x_B, 0_{\neq A})$

$$= \frac{p(x_t \mid x_u, x_B, 0_{\neq A})}{p(0_t \mid x_u, x_B, 0_{\neq A})} p(0_t, x_u, x_B, 0_{\neq A})$$

$$= \frac{p(x_t \mid x_B, 0_{A^c \cup \{u\}})}{p(0_t \mid x_B, 0_{A^c \cup \{u\}})} p(\square)$$

$$= \frac{\prod_{D \subseteq B \cup \{u\}} V_D \prod_{D \subseteq B \cup \{t\}} V_D}{\prod_{D \subseteq B} V_D} = \prod_{D \subsetneq A} V_D \text{ (by induct)}$$

<u>prop.</u> (X, Y) is MRF on $G = G_X \cup G_Y$. with neighbour structure $N_{X \cup Y}$. Then:

i) Marginal dist. of Y is Gibbs dist. on G_Y. with neighbour struc: $\eta_1 \sim \eta_2$ if $\begin{cases} \eta_1 \sim \eta_2 \text{ in } G_Y. \\ \eta_1 \sim x \sim \eta_2.\ x \in G_X. \end{cases}$

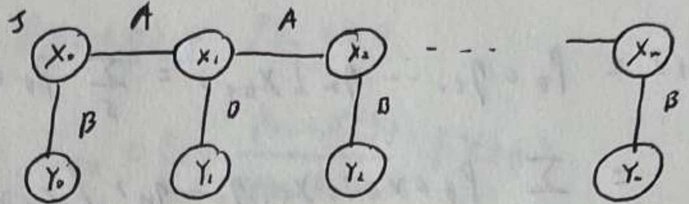
ii) $X \mid Y$ is MRF on G_X. with neighbour struc. N_X.

<u>Pf</u>: i) $p(\eta) = \sum_{S_X} p(x, \eta)$. Written by def.

ii) $p(x \mid \eta) = p(x, \eta)/p(\eta)$.

(2) Hidden Markov Chain:

Consider:



$\theta = (\xi, A, B)$. parameters:

i) ξ is initial dist of X_0.

ii) $A_{ij} = p(X_{t+1}=j \mid X_t = i)$. $B_{ij} = p(Y_t = j \mid X_t = i)$

$A = (A_{ij})_{k \times k}$. $B = (B_{ij})_{k \times v}$. prob. trans. Matrixs.

Rmk: $k=1 \Rightarrow (Y_n)$ i.i.d.

$k = v$. $B = I_k \Rightarrow (X_n)$ is Markov Chain

① Likelihood:

$L(\theta) = P_\theta(y_0 \cdots y_n)$. density of observed data.

$L(\theta) = \sum_{x \in S_X} P_\theta(x, y)$. $|S_X| = k^{n+1}$.

$\Rightarrow$ To calculate $L(\theta)$. We need sum up k^{n+1} times.

Denote: $\alpha_t(x_t) = P_\theta(x_t, y_0, y_1 \cdots y_t)$. $\beta_t(x_t) = P_\theta(y_{t+1} \sim y_n \mid x_t)$

i) Forward Prob.:

$\alpha_0(x_0) = P_\theta(x_0, y_0) = \xi(x_0) B(x_0, y_0)$

$\alpha_{t+1}(x_{t+1}) = P_\theta(x_{t+1}, y_0, \cdots y_{t+1}) = \sum_{x_t} P_\theta(x_t, x_{t+1}, y_0 \cdots y_{t+1})$

$= \sum_{x_t \in S} \alpha_t(x_t) A(x_t, x_{t+1}) B(x_{t+1}, y_{t+1})$

$L_\theta(y) = \sum_{x_n \in S} \alpha_n(x_n)$. obtained by iteratedly calculation

ii) Backward Prob.:

$$\beta_{t-1}(x_{t-1}) = P_\theta(\eta_t, \cdots \eta_n \mid x_{t-1}) = \sum_s P_\theta(x_t, \eta_t \sim \eta_n \mid x_{t-1})$$

$$= \sum_{x_t} P_\theta(x_{t-1}, x_t, \eta_t \sim \eta_n) / \zeta(x_{t-1})$$

$$= \sum_{x_t} A(x_{t-1}, x_t) B(x_t, \eta_t) \beta_t(x_t)$$

$$L(\theta) = \beta_0(x_0) \zeta(x_0) = P_\theta(\eta_1, \eta_2 \cdots \eta_n)$$

② Maximize Likelihood:

After calculating $L(\theta)$ given by $\theta(\zeta, A, B)$.

We want to find $\hat{\theta}$ to max $L(\theta)$. Which is best predictator for dist. of HMM.

Lemma. For $p = (p_i)_1^k$, $q = (q_i)_1^k$ dist. on $(i)_1^k$.

We have: $\sum_1^k p_i \log p_i \geq \sum_1^k p_i \log q_i$.

pf: By $\sum p_i \log q_i/p_i \leq \sum p_i (q_i/p_i - 1) = 0$.

Rmk: Distance between dist. p, q:

$D(p \| q) = \sum p_i \log p_i/q_i$. is called Kullback-Leibler distance.

To maximize $L_\theta = P_\theta(\eta) = \sum_x P_\theta(x, \eta) \iff$ Given $Y = \eta$, maximize $P_\theta(x, \eta)$.

Next, we introduce EM Algorithm:

prop. If $E_{\theta_0}(\log P_{\theta_1}(X,\eta)\mid\eta) > E_{\theta_0}(\log P_{\theta_0}(X,\eta)\mid\eta)$.

Then: $P_{\theta_1}(\eta) > P_{\theta_0}(\eta)$.

Pf. $0 < E_{\theta_0}\left(\log \frac{P_{\theta_1}(X\mid\eta)}{P_{\theta_0}(X\mid\eta)} \mid Y=\eta\right)$

$= \sum_x P_{\theta_0}(x\mid\eta)\log P_{\theta_1}(\eta)/P_{\theta_0}(\eta) - \sum_x P_{\theta_0}(x\mid\eta)\log\frac{P_{\theta_0}(x\mid\eta)}{P_{\theta_1}(x\mid\eta)}$

$= \log P_{\theta_1}(\eta)/P_{\theta_0}(\eta) - \sum\square \leq \log P_{\theta_1}(\eta)/P_{\theta_0}(\eta)$

Note that: $P_\theta(x,\eta) = \xi(x_0)\prod_0^{n-1} A(x_t, x_{t+1})\prod_0^{n} B(x_t,\eta_t)$

$\Rightarrow \log P_\theta(x,\eta) = \log\xi(x_0) + \sum\log A(x_t,x_{t+1}) + \sum\log B(x_t,\eta_t)$

1°) Radomly Choose $\theta_0(\xi_0, A_0, B_0)$

2°) Choose $\theta_1 = (\xi_1, A_1, B_1)$ maximizes $f(\theta) = E_{\theta_0}(\log P_\theta(x,\eta)\mid\eta)$

$= \sum_i P_{\theta_0}(X_0=i\mid\eta)\log\xi(i) + \sum_{t=0}^{n-1}\sum_{i,j} P_{\theta_0}(X_t=i, X_{t+1}=j\mid\eta)\log A(i,j)$

$+ \sum_{t=0}^{n}\sum_i P_{\theta_0}(X_t=i\mid\eta)\log B(i,\eta_t) \overset{\Delta}{=} A_1 + A_2 + A_3.$

For A_1: Choose $\xi_1(i) = P_{\theta_0}(X_0=i\mid\eta)$. by Lemma.

the prob. condition on current data.

For A_2: $\sum_i\sum_j\left(\sum_t^{n-1} P_{\theta_0}(X_t=i, X_{t+1}=j\mid\eta)\right)\log A(i,j))$

Choose $A_1(i,j) = \left(\sum_0^{n-1} P_{\theta_0}(X_t=i, X_{t+1}=j\mid\eta)\right)/\sum_j\left(\sum_0^{n-1}\square\right)$

$\left(\hat{A}(i,j) = \frac{\sum_t I\{X_t=i, X_{t+1}=j\}}{\sum_t I\{X_t=i\}} \approx P(X_t=i\mid X_{t+1}=j)\right)$

For A_3: By Lemma analgously, Choose:

$B_1(i,j) = \sum_{t:\eta_t=j} P_{\theta_0}(X_t=i\mid\eta) / \sum_j\sum_{t:\eta_t=j} P_{\theta_0}(X_t=i\mid\eta)$

$\left(\hat{B}(i,j) = \frac{\sum_t^n I\{X_t=i, Y_t=j\}}{\sum_t^n I\{X_t=i\}}. \approx P(Y_t=i\mid X_t=j)\right)$

3°) Calculate $\theta_1 = (\xi_1, A_1, B_1)$

Consider $\gamma_t(i,j) = P_{\theta_0}(X_t = i, X_{t+1} = j \mid \eta)$ which

can express θ_1. $\Leftarrow$ $P_{\theta_0}(X_t, X_{t+1}, \eta)$. Since $P_{\theta_0}(\eta)$

can be calculated by forward prob.

$$P_{\theta_0}(X_t, X_{t+1}, \eta) = P_{\theta_0}(\eta_0^t, X_t, X_{t+1}, \eta_{t+1}^n)$$

$$= \alpha_t(x_t)\, A_0(x_t, x_{t+1})\, P_{\theta_0}(\eta_{t+1}^n \mid X_{t+1})$$

$$P_{\theta_0}(\eta_{t+1}^n \mid X_{t+1}) = P_{\theta_0}(\eta_{t+1}, \eta_{t+2}^n \mid X_{t+1})$$

$$= B_0(X_{t+1}, \eta_{t+1})\, \beta_{t+1}(X_{t+1})$$

$$\Rightarrow \gamma_t(i,j) = \alpha_t(x_t)\, A_0(x_t, x_{t+1})\, B_0(x_{t+1}, \eta_{t+1})\, \beta_{t+1}(x_t) \Big/ \square$$

$$\square = \sum_{i,j} \alpha_t(i)\, A_0(i,j)\, B(j, \eta_{t+1})\, \beta_{t+1}(j).$$

Def: $\xi_1(i) = \sum_j \gamma_0(i,j)$

$$A_1(i,j) = \sum_{t=0}^{n-1} \gamma_t(i,j) \Big/ \sum_l \sum_{t=0}^{n-1} \gamma_t(i,l)$$

$$B_1(i,j) = \frac{\sum_{t=0}^{n-1} \sum_l \gamma_t(i,l)\, I_{\{\eta_t = j\}} + \sum_m \gamma_{n-1}(m,i)\, I_{\{\eta_n = j\}}}{\sum_{t=0}^{n-1} \sum_l \gamma_t(i,l) + \sum_m \gamma_{n-1}(m,i)}$$

4°) Replace θ_0 by θ_1. Repeat the procedure.