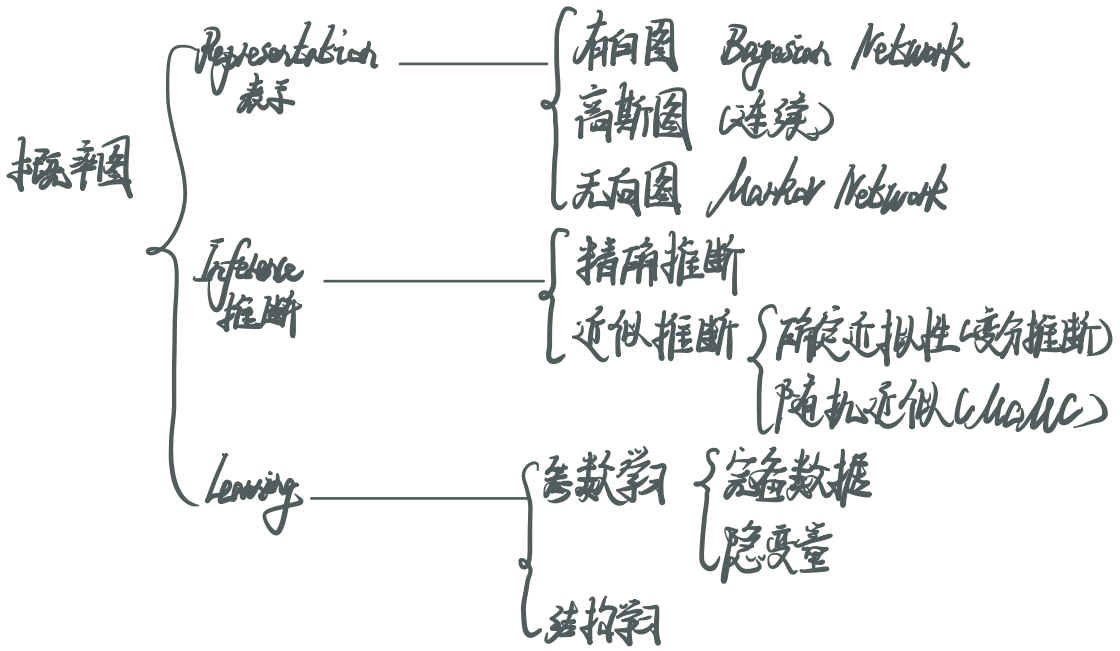


1. Basics

高维随机变量 { 边缘概率 $P(x_i)$; 条件概率 $P(x_i|x_j)$ }

{ Sum Rule: $P(x_1)=\int P(x_1,x_2)dx_2$

Product Rule: $P(x_1,x_2)=P(x_2)\cdot P(x_1|x_2)=P(x_2)\cdot P(x_1|x_2)$

Chain Rule: $P(x_1,x_2\cdots x_p)=\prod_{i=1}^{p} P(x_i|x_1,x_2\cdots x_{i-1})\cdot P(x_1)$

eg: $P(A_1,A_2,A_3)=P(A_3|A_2,A_1)\cdot P(A_2,A_1)=P(A_3|A_2,A_1)\cdot P(A_2|A_1)\cdot P(A_1)$

Bayesian Rule: $P(x_2|x_1)=\frac{P(x_1,x_2)}{P(x_1)}=\frac{P(x_1,x_2)}{\int P(x_1,x_2)dx_2}=\frac{P(x_2)\cdot P(x_1|x_2)}{\int P(x_2)\cdot P(x_1|x_2)dx}$

困境:维度高,计算复杂 $P(x_1,x_2\cdots x_p)$ 计算量大

↓

简化 → 相互独立 → $P(x_1\cdots x_p)=\prod_{i=1}^{p} P(x_i)$ (Chain Rule → $P(x|y)=\prod_{i=1}^{p} P(x_i|y)$)

简化 → Markov Property → $x_{i+1}\perp x_{i-1}|x_i$ 子<i

↳ HMM (未发Markov假设)

x_1 x_2 … x_p

条件独立性 → $x_A\perp x_B|x_C$

Bayesian Network

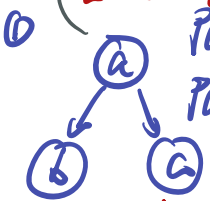
$P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i \mid x_1, x_2, \cdots, x_{i-1}) \cdot P(x_1)$

条件独立性：$x_A \perp x_B \mid x_C$

因子分解：$P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i \mid x_{pa(i)})$ ← 父节点集合

tail head

tail to tail



因子分解

$P(a,b,c) = P(a) \cdot P(b|a) \cdot P(c|a)$

$P(a,b,c) = P(a) \cdot P(b|a) \cdot P(c|b,a)$

chain

$\Rightarrow P(c|a) = P(c|b,a) \rightarrow (c \perp b) \mid a$

$P(c|a) \cdot P(b|a) = P(c|b,a) \cdot P(b|a) = P(b,c|a)$

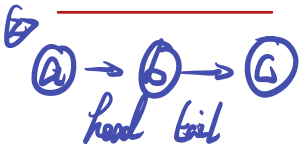
$P(c|a) \cdot P(b|a) = P(b,c|a)$

$(c \perp b) \mid a$

⇑

若 a 被观测，则路径被阻塞

head to tail



$(a \perp c) \mid b$

⇑ 独立

b 被观测，则路径被阻塞

$P(a,b,c) = P(c) \cdot P(b|c) \cdot P(a|b)$

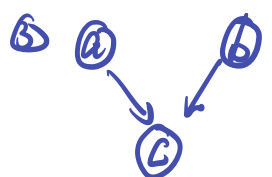
$= P(c) \cdot P(b|c) \cdot P(a|b,c)$

$P(a|b) = P(a|b,c)$

$P(a|b) \cdot P(c|b) = P(a|b,c) \cdot P(c|b) = P(a,c|b)$

=

head to head



默认 a⊥b，路径堵塞

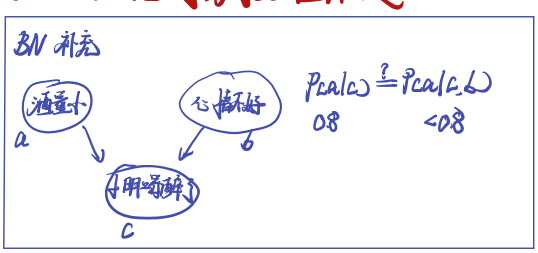
若 c 被观测，则路径能通

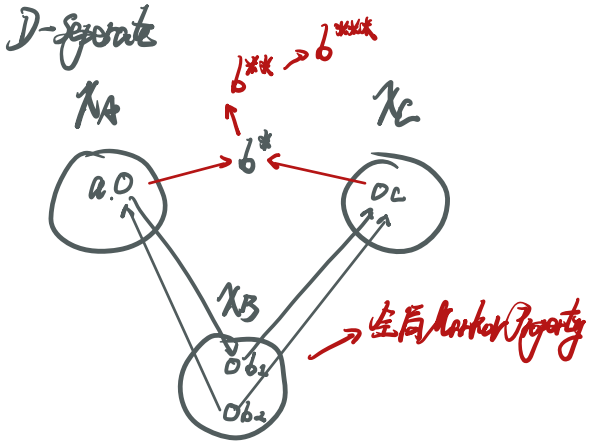
$P(a,b,c) = P(c|a,b) \cdot P(a) \cdot P(b)$

$= P(c|a,b) \cdot P(b|a) \cdot P(a)$

$\Rightarrow P(b) = P(b|a)$

$a \perp b$



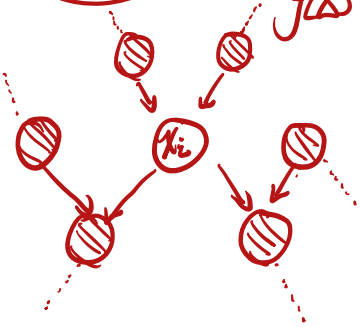


$$P(x_i|x_{-i}) = \frac{P(x_i, x_{-i})}{P(x_{-i})} = \frac{P(x)}{\int_{x_i} P(x)\,dx_i} \qquad \int_{x_i} P(x_{-i})\,P(x_i|x_{pa(i)})$$

x_i 存在 Δ / x_i 无关 Δ

$$= \prod_{j=1}^{p} P(x_{ij}|x_{pa(ij)}) \qquad \int_{x_i} \prod_{j=1}^{p} P(x_{ij}|x_{pa(ij)})\,dx_i$$

$$f(\Delta) = \begin{cases} P(x_i|x_{pa(i)}) \\ P(x_{child(i)}|x_i, x_{pa(child(i))}) \end{cases}$$



Ex:

Bayesian Network

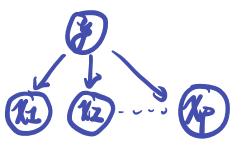
从单一到混合

从有限到无限

① 空间 (离散→连续)

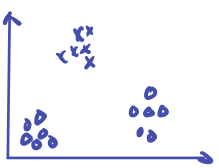
② 时间

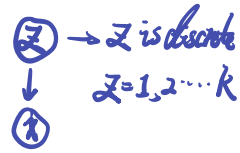
$(x_1 \perp x_2)|z$



单一: Naive Bayes → $P(x|y) = \prod_{i=1}^{p} P(x_i|y)$

混合: GMM





时间: Markov Chain

Gaussian Process (无限维高斯分布)

→ 动态模型

HMM (离散) LDS Particle Filter

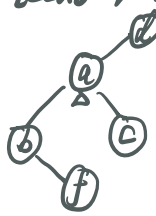
连续: Gaussian Bayes Network

Markov Network

① $X_A \perp X_C \mid X_B$

Global Markov

② Local Markov



$a \perp$ 集合-a-邻域|邻居

$a \perp \{e, f\} \mid \{b, c, d\}$

③成对 Markov

$X_i \perp X_j \mid X_{-i-j}$（无向）

条件独立性体现在

三个方面：① ② ③

① ⇔ ② ⇔ ③

因子分解

Clique

图：团，最大团

↳一个关于结点的集合，集合中结点相互都连通

$$P(X) = \frac{1}{Z}\prod_{i=1}^{k}\phi(X_{ci}) = \frac{1}{Z}\prod_{Q\in C}\psi_Q(X_Q)$$

$$Z = \sum_X \prod_{i=1}^{k}\phi(X_{ci}) = \sum_X \prod_{Q\in C}\psi_Q(X_Q)$$

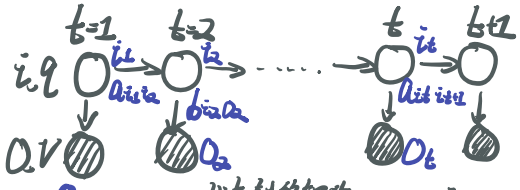
$$P(X) = \frac{1}{Z}\prod_{i=1}^{k}\psi(X_{ci})$$

C_i：最大团，X_{ci}：最大团随机变量集合

$\psi(X_{ci})$：势函数，必须为 + ⟶ $\psi(X_{ci}) = \exp\{-E(X_{ci})\}$

$Z = \sum_X \prod_{i=1}^{k}\psi(X_{ci}) = \sum_{x_1}\sum_{x_2}\cdots\sum_{x_p}\prod_{i=1}^{k}\psi(X_{ci})$ ↳ $P(X)$ Gibbs Distribution

Hidden Markov Model (HMM)



$A = [a_{ij}]$ 状态转移矩阵

i. $\lambda = (\pi, A, B)$ → 发射矩阵

初始 pro dist

$i_t = \{q_1, q_2 \cdots q_N\}$

$\pi = \{p_1, p_2 \cdots p_N\}$

值域

观测变量 $O, O_1, O_2 \cdots O_t \rightarrow V = \{v_1, v_2 \cdots v_M\}$

状态变量 $i, i_1, i_2 \cdots i_t \rightarrow Q = \{q_1, q_2 \cdots q_N\}$

$A = [a_{ij}]$, $a_{ij} = P(i_{t+1} = q_j \mid i_t = q_i)$

$B = [b_{jk}]$, $b_{jk} = P(O_t = v_k \mid i_t = q_j)$

ii 两个假设：①齐次 Markov ②观测独立

↓ 无后效性

① $P(i_{t+1} \mid i_t, \cdots i_1, O_t, O_{t-1}, \cdots O_1) = P(i_{t+1} \mid i_t)$

② $P(O_t \mid i_t, i_{t-1}, \cdots i_1, O_{t-1} \cdots O_1) = P(O_t \mid i_t)$

iii 三个问题

① Evaluation → $P(O \mid \lambda)$ → 前后向

② Learning → λ 如何求？ → EM

Baum Welch

③ Decoding → $\arg\max_i P(I \mid O)$

↳ {预测 → $P(i_{t+1} \mid O_1, O_2 \cdots O_t)$

滤波 → $P(i_t \mid O_1, O_2 \cdots O_t)$

Evaluation: Given λ 求 $P(O \mid \lambda)$

$$P(O \mid \lambda) = \sum_I P(I, O \mid \lambda) = \sum_I P(O \mid I, \lambda) \cdot P(I \mid \lambda)$$

$$P(I \mid \lambda) = P(i_1 \cdots i_T \mid \lambda) = \underbrace{P(i_T \mid i_1 \cdots i_{T-1}, \lambda)}_{P(i_T \mid i_{T-1})} \cdot P(i_1 \cdots i_{T-1} \mid \lambda) = a_{i_{T-1} i_T} \cdot a_{i_{T-2} i_{T-1}} \cdots a_{i_1 i_2} \cdot \pi(i_1) = \pi(i_1) \prod_{t=2}^{T} a_{i_{t-1} i_t}$$

$a_{i_{T-1} i_T}$

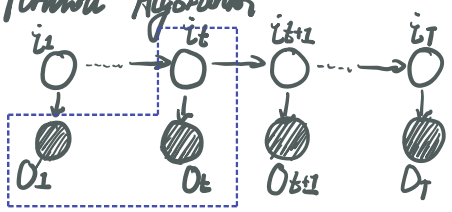
$$P(O \mid I, \lambda) = \prod_{t=1}^{T} b_{i_t}(O_t)$$

$b_{i_1}(O_1)\, b_{i_2}(O_2) \cdots b_{i_T}(O_T)$

$O(N^T)$

$$\therefore P(O \mid \lambda) = \sum_I \pi(i_1) \prod_{t=2}^{T} a_{i_{t-1} i_t} \prod_{t=1}^{T} b_{i_t}(O_t) = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_T} \pi(i_1) \prod_{t=2}^{T} a_{i_{t-1} i_t} \prod_{t=1}^{T} b_{i_t}(O_t)$$

Forward Algorithm



记: $\alpha_t(i) = P(O_1 \cdots O_t, i_t = q_i \mid \lambda)$

$\alpha_T(i) = P(O, i_T = q_i \mid \lambda)$

$$P(O \mid \lambda) = \sum_{i=1}^{N} P(O, i_t = q_i \mid \lambda) = \sum_{i=1}^{N} \alpha_T(i)$$

$$\alpha_{t+1}(j) = P(O_1 \cdots O_t, O_{t+1}, i_{t+1} = q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_1 \cdots O_t, O_{t+1}, i_{t+1} = q_j, i_t = q_i \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_{t+1} \mid O_1 \cdots O_t, i_t = q_i, i_{t+1} = q_j, \lambda) \cdot P(O_1 \cdots O_t, i_t = q_i, i_{t+1} = q_j \mid \lambda)$$

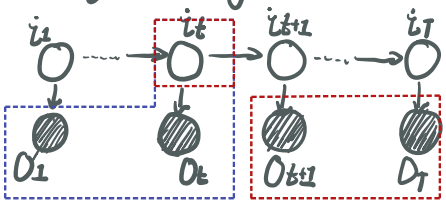
$$= \sum_{i=1}^{N} P(O_{t+1} \mid i_{t+1} = q_j) \cdot P(O_1 \cdots O_t, i_t = q_i, i_{t+1} = q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_{t+1} \mid i_{t+1} = q_j) \cdot P(i_{t+1} = q_j \mid O_1 \cdots O_t, i_t = q_i, \lambda) \cdot P(O_1 \cdots O_t, i_t = q_i \mid \lambda)$$

$$= \sum_{i=1}^{N} P(O_{t+1} \mid i_{t+1} = q_j) \cdot P(i_{t+1} = q_j \mid i_t = q_i) \cdot \alpha_t(i)$$

$$= b_j(O_{t+1}) \cdot a_{ij} \cdot \alpha_t(i)$$

Backforward Algorithm



记 $\beta_t(i) = P(O_{t+1} \cdots O_T \mid i_t = q_i, \lambda)$

$\beta_1(i) = P(O_2, \cdots O_T \mid i_t = q_i, \lambda)$

$$P(O \mid \lambda) = P(O_1, \cdots O_T \mid \lambda)$$

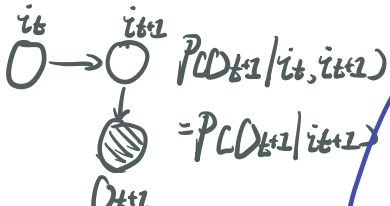
$$= \sum_{i=1}^{N} P(O_1, \cdots O_T, i_1 = q_i)$$

$$= \sum_{i=1}^{N} P(O_1 \cdots O_T \mid i_1 = q_i) \cdot P(i_1 = q_i)$$

$$= \sum_{i=1}^{N} P(O_1 \mid O_2 \cdots O_T, i_1 = q_i) \cdot P(O_2 \cdots O_T \mid i_1 = q_i) \cdot \pi_i$$

$$= \sum_{i=1}^{N} P(O_1 \mid i_1 = q_1) \cdot \beta_1(i) \cdot \pi_i$$

$$= \sum_{i=1}^{N} b_i(O_1) \pi_i \beta_1(i)$$



$P(O_{t+1} \mid i_t, i_{t+1}) = P(O_{t+1} \mid i_{t+1})$

$$\beta_t(i) = P(O_{t+1} \cdots O_T \mid i_t = q_i, \lambda)$$

$$= \sum_{j=1}^{N} P(O_{t+1} \cdots O_T, i_{t+1} = q_j \mid i_t = q_i)$$

$$= \sum_{j=1}^{N} P(O_{t+1}, \cdots O_T \mid i_{t+1} = q_j, i_t = q_i) \cdot P(i_{t+1} = q_j \mid i_t = q_i)$$

$$= \sum_{j=1}^{N} P(O_{t+1} \cdots O_T \mid i_{t+1} = q_j) \cdot a_{ij}$$

$$= \sum_{j=1}^{N} P(O_{t+1} \mid O_{t+2} \cdots O_T, i_{t+1} = q_j) \cdot P(O_{t+2} \cdots O_T \mid i_{t+1} = q_j) \cdot a_{ij}$$

$$= \sum_{j=1}^{N} P(O_{t+1} \mid i_{t+1} = q_j) \cdot \beta_{t+1}(j) \cdot a_{ij}$$

$$= \sum_{j=1}^{N} b_j(O_{t+1}) \cdot a_{ij} \cdot \beta_{t+1}(j)$$

learning

$\lambda_{MLE} = \arg\max_{\lambda} P(O|\lambda)$ → Baum-Welch EM

$\theta^{(t+1)} = \arg\max_{\theta} \int_Z \log P(X,Z|\theta) \cdot P(Z|X,\theta^{(t)}) dZ$

X: 观测 → O　Z: 隐变量 → I → 离散　θ: 参数 → λ

$\lambda^{(t+1)} = \arg\max_{\lambda} \sum_I \log P(O,I|\lambda) \cdot P(I|O,\lambda^{(t)})$

$P(I|O,\lambda^{(t)}) = \frac{P(I,O|\lambda^{(t)})}{P(O|\lambda^{(t)})}$ → constant

$\lambda^{(t+1)} = \arg\max_{\lambda} \sum_I \log P(O,I|\lambda) \cdot P(I,O|\lambda^{(t)})$

$\lambda^{(t)} = (\pi^{(t)}, A^{(t)}, B^{(t)})$

$P(O|\lambda) = \sum_I P(O,I|\lambda) = \sum_{i_1} \sum_{i_2} \cdots \sum_{i_T} \pi_{i_1} a_{i_1 i_2} \cdots a_{i_{T-1} i_T} b_{i_1}(o_1) \cdots b_{i_T}(o_T)$

$Q(\lambda, \lambda^{(t)}) = \sum_I \log P(O,I|\lambda) \cdot P(O,I|\lambda^{(t)})$

$= \sum_I [(\log \pi_{i_1} + \sum_{t=2}^{T} \log a_{i_{t-1} i_t} + \sum_{t=1}^{T} b_{i_t}(o_t)] \cdot P(O,I|\lambda^{(t)})]$

$\pi^{(t+1)} = \arg\max_{\pi} Q(\lambda, \lambda^{(t)})$

$= \arg\max_{\pi} \sum_I [\log \pi_{i_1} \cdot P(O,I|\lambda^{(t)})]$

$= \arg\max_{\pi} \sum_{i_1}^{N} \sum_{i_2} \cdots \sum_{i_T} [\log \pi_{i_1} \cdot P(O, i_1 \cdots i_T|\lambda^{(t)})]$ → 将 $\sum_{i_2} \cdots \sum_{i_T} P(O, i_1 \cdots i_T|\lambda^{(t)})$ 消掉

$= \arg\max_{\pi} \sum_{i_1}^{N} [\log \pi_{i_1} \cdot P(O, i_1 = q_{i_1}|\lambda^{(t)})]$　s.t. $\sum_{i_1}^{N} \pi_{i_1} = 1$

$L(\pi, \eta) = \sum_{i_1}^{N} [\log \pi_{i_1} \cdot P(O, i_1 = q_{i_1}|\lambda^{(t)})] + \eta(\sum_{i_1}^{N} \pi_{i_1} - 1)$

$\frac{\partial L}{\partial \pi_{i_1}} = \frac{1}{\pi_{i_1}} P(O, i_1 = q_{i_1}|\lambda^{(t)}) + \eta = 0$

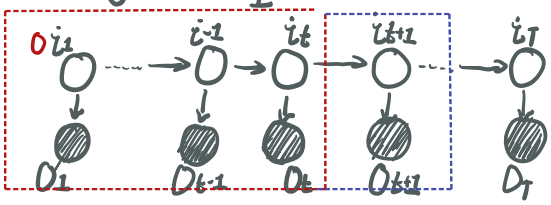
$\sum_{i_1}^{N} [P(O, i_1 = q_{i_1}|\lambda^{(t)}) + \pi_{i_1} \eta] = 0$

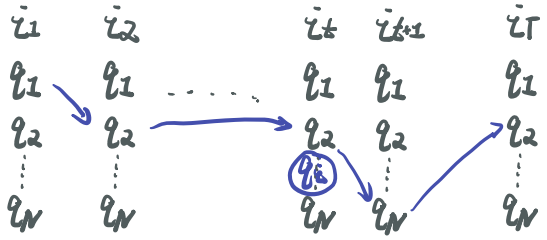
$P(O|\lambda^{(t)}) + \eta = 0$

$\therefore \eta = -P(O|\lambda^{(t)})$ ⟹ $\pi_{i_1}^{(t+1)} = \frac{P(O, i_1 = q_{i_1}|\lambda^{(t)})}{P(O|\lambda^{(t)})}$

$\pi^{(t+1)} = (\pi_1^{(t+1)} \cdots \pi_N^{(t+1)})$

Decoding : $\hat{I} = \arg\max_{I} P(I|O,\lambda)$ → Viterbi

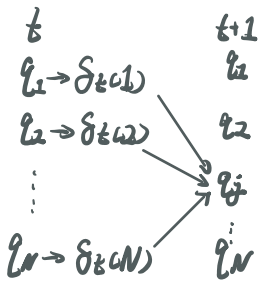




$$\delta_t(i) = \max_{i_1, i_2 \cdots i_{t-1}} P(O_1, O_2 \cdots O_t, i_1, i_2 \cdots i_{t-1}, i_t = q_i)$$

$$\delta_{t+1}(j) = \max_{i_1, i_2 \cdots i_t} P(O_1, O_2 \cdots O_t, O_{t+1}, i_1, i_2 \cdots i_{t-1} = q_j)$$

$$= \max_{1 \le i \le N} \delta_t(i) \cdot a_{ij} \cdot b_j(O_{t+1})$$



$$\psi_{t+1}(j) = \arg\max_{1 \le i \le N} \delta_t(i) \cdot a_{ij}$$

Markov Chain & Monte Carlo (MCMC)

Monte Carlo Method: 基于采样的随机近似法

$P(Z|X)$ ← X: observed data; Z: latent data

$$\longrightarrow E_{Z|X}[f(Z)] = \int P(Z|X) \cdot f(Z)\, dZ \approx \frac{1}{N}\sum_{i=1}^{N} f(Z_i)$$

Variational Inference

$D=\{(x_i,y_i)\}_{i=1}^{N}$ $x\in\mathbb{R}^p$

频率角度 → 优化问题 {
回归 {
$f(w)=w^Tx$ Loss Function: $L(w)=\sum_{i=1}^{N}\|w^Tx_i-y_i\|^2$
Model Strategy $\hat{w}=\arg\min_{w\in\mathbb{R}^p} L(w)$
解法 Algo {
①解析解: $\frac{\partial L}{\partial w}=0\Rightarrow w^*=(X^TX)^{-1}XY$
②数值解: GD/SGD
}
SVM {
①$f(w)=sign(w^Tx+b)$
②Loss Function: $\min\frac{1}{2}w^Tw$, s.t. $y_i w^Tx\geq 1$, $i=1,\dots,N$ (硬间隔)
③QP / Lagrange 对偶
}
}

ML: $\hat{\theta}=\arg\max_\theta \log P(X|\theta)$

$\theta^{(t+1)}=\arg\max_\theta \int \log P(X,Z|\theta)\cdot P(Z|X,\theta^{(t)})\,dZ$

贝叶斯角度 → 积分问题 $P(\theta|X)=\frac{P(X|\theta)\cdot P(\theta)}{P(X)}=\frac{P(X|\theta)\cdot P(\theta)}{\int_\theta P(X|\theta)\cdot P(\theta)\,d\theta}$

(likelihood, prior, posterior, $P(X)$)

~ Decision $X\rightarrow$ N个样本

$\hat{x}$, $P(\bar{x}|X)=\int_\theta P(\bar{x},\theta|X)\,d\theta=\int_\theta P(\bar{x}|\theta)\cdot P(\theta|X)\,d\theta$ (posterior)

$=E_{\theta|X}[P(\bar{x}|\theta)]$

贝叶斯 Inference {
精确推断
近似推断 {
确定性近似 → VI
随机近似 → MCMC, MH, Gibbs
}
}

X: Observed Data

Z: Latent Variable + parameter

(X, Z): Complete Data

VI (mean field) → classical VI

Assumption: $q(Z) = \prod_{i=1}^{M} q_i(Z_i)$ → independent

$\log q_j(Z_j) = E_{\prod_{i \neq j} q_i(Z_i)}[\log p(X, Z|\theta)] + C$

$$\log p(X) = \log p(X, Z) - \log p(Z|X)$$

$$= \log \frac{p(X,Z)}{q(Z)} - \log \frac{p(Z|X)}{q(Z)}$$

两边同时取期望

左边: $\int_Z \log p(X) q(Z) dZ$

$= \log p(X) \underbrace{\int_Z q(Z) dZ}_{1}$

$= \log p(X)$

左边: $\underbrace{\int_Z q(Z) \log \frac{p(X,Z)}{q(Z)} dZ}_{\text{ELBO evidence lower Bound}} - \underbrace{\int_Z q(Z) \log \frac{p(Z|X)}{q(Z)}}_{KL(q\|p)}$

$= \mathcal{L}(q) + \underbrace{KL(q\|p)}_{\geq 0}$

$\mathcal{L}(q)$ 变分

$$\hat{q}(Z) = \arg\max_{q(Z)} \mathcal{L}(q) \Rightarrow \hat{q}(Z) \approx p(Z|X)$$

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

$$\mathcal{L}(q) = \underbrace{\int_Z q(Z) \log p(X, Z) dZ}_{①} - \underbrace{\int q(Z) \log q(Z) dZ}_{②}$$

$$① = \int_Z \prod_{i=1}^{M} q_i(Z_i) \log p(X, Z) dZ_1 dZ_2 \cdots dZ_M$$

$$= \int_{Z_j} q_j(Z_j) \left(\int_{\prod_{i \neq j} Z_i} \prod_{i \neq j} q_i(Z_i) \log p(X, Z) dZ_1 dZ_2 \cdots dZ_M \right) dZ_j$$

$$= \int_{Z_j} q_j(Z_j) \left(\underbrace{\int_{\prod_{i \neq j} Z_i} \log p(X, Z) \prod_{i \neq j} q_i(Z_i) dZ_i}_{E_{\prod_{i \neq j} q_i(Z_i)}[\log p(X, Z)] = \log \tilde{p}(X, Z_j)} \right)$$

$$= \int_{Z_j} q_j(Z_j) \cdot E_{\prod_{i \neq j} q_i(Z_i)}[\log p(X, Z)] \cdot dZ_j$$

$$② = \int_Z q(Z) \log q(Z) dZ$$

$$= \int_Z \prod_{i=1}^{M} q_i(Z_i) \cdot \sum_{i=1}^{M} \log q_i(Z_i) dZ$$

$$= \int_Z \prod_{i=1}^{M} q_i(Z_i) \left[\log q_1(Z_1) + \log q_2(Z_2) \cdots \log q_M(Z_M) \right] dZ$$

$$= \sum_{i=1}^{M} \int q_i(Z_i) \log q_i(Z_i) dZ_i$$

$$= \int_{Z_j} q_j(Z_j) \log q_j(Z_j) dZ_j + C$$

$$\int_Z \prod_{i=1}^{M} q_i(Z_i) \cdot \log q_1(Z_1) dZ$$

$$= \int_Z q_1 q_2 \cdots q_M \cdot \log q_1 dZ$$

$$= \int_{Z_1 \cdots Z_M} q_1 \cdots q_M \log q_1 dZ_1 \cdots dZ_M$$

$$= \int_{Z_1} q_1 \log q_1 dZ_1 \cdot \underbrace{\int_{Z_2} q_2 dZ_2}_{1} \cdots \underbrace{\int_{Z_M} q_M dZ_M}_{1}$$

$$= \int_{Z_1} q_1 \log q_1 dZ_1$$

$$① - ② = \int_{z_j} q_j(z_j) \cdot \log \frac{\hat{p}(x, z_j)}{q_j(z_j)} dz_j + C$$
$$= -KL(q_j \| \hat{p}(x, z_j)) \leq 0$$

∴ $q_j(z_j) = \hat{p}(x, z_j)$