Hidden Markov Model

频率派 → 统计机器学习

↓ 优化

① model: $f(w)=w^Tx+b$

② strategy: loss function

③ algorithem: SGD、GD、牛顿法、拟牛顿法

贝叶斯派 → 概率图模型

↓

Inference → $P(Z|X)$ → 积分问题

↓ 数值积分 MCMC

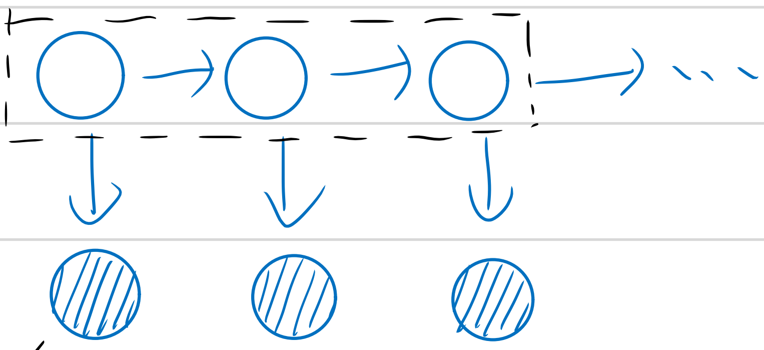
概率图
- 有向 — Bayesian Network
- 无向 — Markov Random Field (Markov Network)

+time → Dynamic Model
- HMM
- Kalman Filter
- Particle Filter

↳
- time
- mixture

System state

隐变量

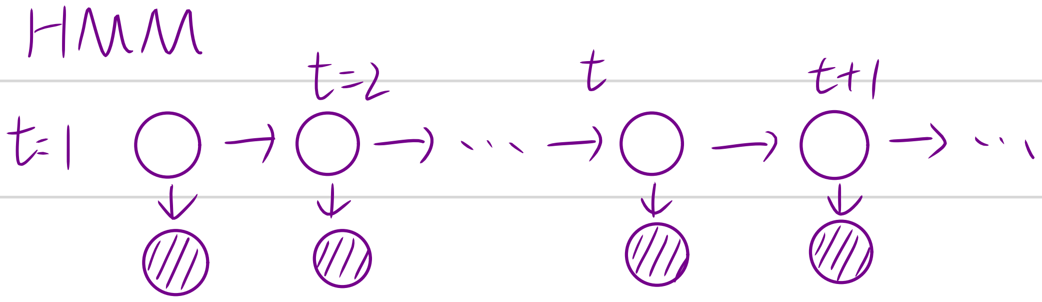


观测变量

state → 离散 HMM

连续
- 线性 → Kalman Filter
- 非线性 → Particle Filter

HMM



$\lambda = (\pi, A, B)$

- A → 状态转移矩阵
- B → 发射矩阵
- π → 初始prob dist

观测变量 o，$o_1, o_2, \cdots o_t, \cdots \rightarrow V=\{v_1, v_2, \cdots v_M\}$

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

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

$B=[b_j(k)]$，$b_j(k)=P(o_t=v_k \mid i_t=q_j)$

$t=1$ 时刻，$i_1=q_1, q_2, \cdots q_N \Rightarrow \pi=(P_1, P_2, \cdots P_N)$

根据初始概率分布生成第一个状态

两个假设：①齐次Markov假设。②观测独立假设。

$P(i_{t+1} \mid i_t, i_{t-1}, \cdots i_1, o_t, o_{t-1}, \cdots o_1) = P(i_{t+1} \mid i_t)$

$P(o_t \mid i_t, i_{t-1}, \cdots i_1, o_{t-1}, \cdots, o_1) = P(o_t \mid i_t)$

iii 三个问题

① Evaluation $\rightarrow P(O \mid \lambda) \rightarrow$ 前向后向

② Learning $\rightarrow \lambda$ 如何求 $\rightarrow$ EM Baum Welch $\lambda = \arg\max P(O \mid \lambda)$

③ Decoding $\rightarrow \arg\max P(I \mid O)$ $I=\arg\max P(I \mid O)$

$\rightarrow$ 预测 $\rightarrow P(i_{t+1} \mid o_1, o_2, \cdots o_t)$ 预测问题

滤波 $\rightarrow P(i_t \mid o_1, o_2, \cdots o_t)$

总结：一个模型，两个假设，三个问题

前向算法

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

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

$P(I \mid \lambda) = P(i_1, i_2, \ldots i_T \mid \lambda) = P(i_T \mid i_1, i_2, \cdots i_{T-1}, \lambda) \cdot P(i_1, i_2, \cdots i_{T-1}, \lambda)$

$= a_{i_{T-1} i_T} \, a_{i_{T-2} i_{T-1}} \cdot a_{i_{T-3} i_{T-2}} \cdots \pi(i_1)$

$= \pi(a_{i_1}) \cdot \prod_{t=2}^{T} a_{i_{t-1} i_t}$

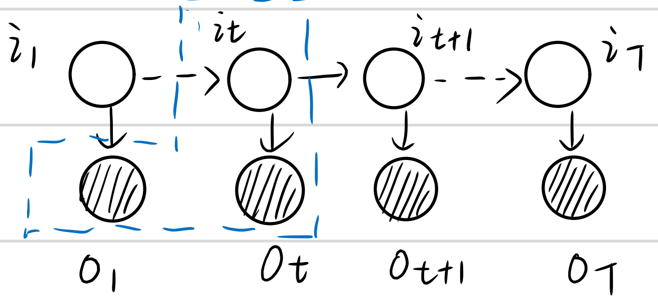
$P(O \mid I, \lambda) = \prod_{t=1}^{T} b_{i_t}(o_t)$

$P(O \mid \lambda) = \sum_I \pi(a_{i_1}) \prod_{t=2}^{T} a_{i_{t-1}, i_t} \prod_{t=1}^{T} b_{i_t}(o_t)$

$$= \underbrace{\sum_{i_1}\sum_{i_2}\cdots\sum_{i_T}}_{N} \pi(a_{i_1})\prod_{t=2}^{T} a_{i_{t-1}i_t}\prod_{t=1}^{T} b_{i_t}(O_t)$$

$O(N^T)$ 指数级增长

Forward Algorithm



记 $\boxed{\alpha_t(i)=P(O_1,\ldots,O_t,i_t=q_i|\lambda)}$

$\alpha_T(i)=P(O,i_t=q_i|\lambda)$

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

$$=\sum_{i=1}^{N}\alpha_T(i)$$

$\alpha_{t+1}(j)=P(O_1,\cdots,O_t,O_{t+1},i_{t+1}=q_j|\lambda)$ 对联合分布在某一变量求积分得到边缘分布

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

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

根据观测独立假设 $=\sum_{i=1}^{N}P(O_{t+1}|O_t)\cdot P(O_1,\cdots,O_t,i_t=q_i,i_{t+1}=q_j|\lambda)$

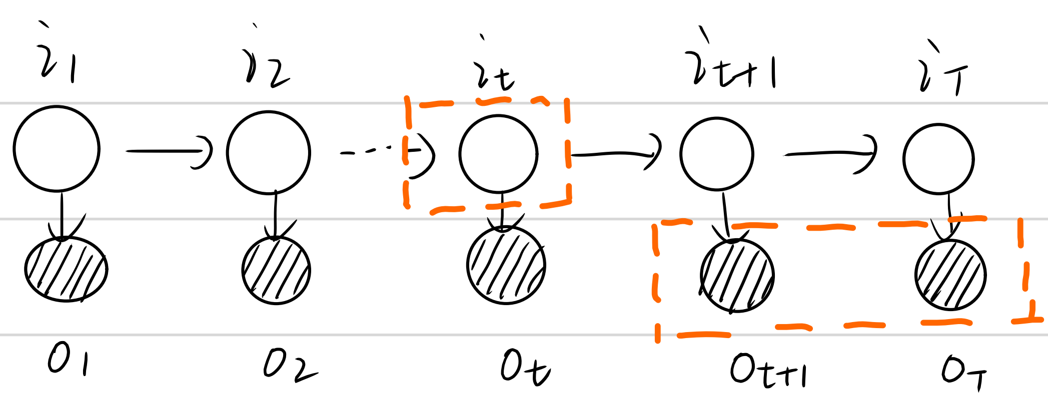
$$=\sum_{i=1}^{N}P(O_{t+1}|i_{t+1}=q_j)\cdot \underline{P(i_{t+1}=q_j|O_1,\cdots,O_t,i_t=q_i,\lambda)}\,\boxed{P(O_1,\cdots,O_t,i_t=q_i|\lambda)}$$

→ 根据齐次马尔可夫假设 $P(i_{t+1}=q_j|i_t=q_i,\lambda)$

$$=\sum_{i=1}^{N}\underbrace{P(O_{t+1}|i_{t+1}=q_j)}_{b_j(O_{t+1})}\cdot\underbrace{P(i_{t+1}=q_j|i_t=q_i,\lambda)}_{a_{ij}}\,\alpha_t(i)$$

$$=\sum_{i=1}^{N}b_j(O_{t+1})\,a_{ij}\,\alpha_t(i)$$

后向算法



记 $\beta_t(i)=P(o_{t+1},\cdots o_T|i_t=q_i,\lambda)$

$\vdots$

$\beta_1(i)=P(o_2,\cdots o_T|i_1=q_i,\lambda)$

$$P(O|\lambda)=P(o_1,\cdots,o_T|\lambda)$$

$$=\sum_{i=1}^{N}P(o_1,\cdots o_T,i_1=q_i)$$

$$=\sum_{i=1}^{N}P(o_1,\cdots,o_T|i_1=q_i)\cdot\boxed{P(i_1=q_i)}\ \pi_i$$

$$=\sum_{i=1}^{N}\underbrace{P(o_1|o_2,\cdots o_T,i_1=q_i)}_{\text{观测独立假设}}\cdot P(o_2,\cdots o_T|i_1=q_i)\cdot\pi_i$$

$$=\sum_{i=1}^{N}P(o_1|i_1=q_i)\,\beta_1(i)\cdot\pi_i$$

$$=\sum_{i=1}^{N}b_i(o_1)\pi_i\beta_1(i)$$

$$\beta_t(i)=P(o_{t+1},\cdots o_T|i_t=q_i)$$

$$=\sum_{j=1}^{N}P(o_{t+1},\cdots o_T,i_{t+1}=q_j|i_t=q_i)$$

$$=\sum_{j=1}^{N}\underline{P(o_{t+1},\cdots o_T|i_{t+1}=q_j,i_t=q_i)}\cdot\underbrace{P(i_{t+1}=q_j|i_t=q_i)}_{a_{ij}}=\sum_{j=1}^{N}P(o_{t+1},\cdots,o_T|i_{t+1}=q_j)\cdot a_{ij}$$

$$=\sum_{j=1}^{N}\underbrace{P(o_{t+1}|o_{t+2},\cdots o_T,i_{t+1}=q_j)}_{P(o_{t+1}|i_{t+1}=q_j)}\cdot\underbrace{P(o_{t+2},\cdots o_T|i_{t+1}=q_j)}_{\beta_{t+1}(j)}\cdot a_{ij}$$

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

Baum-Welch算法

EM算法：$\theta^{(t+1)}=\arg\max_{\theta}\int_z\log P(x,z|\theta)\cdot P(z|x,\theta^{(t)})\,dz$

x：观测 $\to O$ $\quad z$：隐变量 $\to I$ $\quad \theta$：参数 $\to\lambda$

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

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

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

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

由之前推算知：$P(O|\lambda)=\sum_I P(O,I|\lambda)=\sum_{i_1}\cdots\sum_{i_T}\pi_{i_1}\cdot\prod_{t=2}^{T}a_{i_{t-1},i_t}\cdot\prod_{t=1}^{T}b_{i_t}(o_t)$

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

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

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

$$= \arg\max_{\pi} \sum_{i_1} \cdots \sum_{i_T} [\log \pi_{i_1} \cdot P(O, i_1, \cdots, i_T \mid \lambda^{(t)})]$$

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

$$= \arg\max_{\pi} \sum_{i=1}^{N} [\log \pi_i \, P(O, i_1 = q_i \mid \lambda^{(t)})] \quad (s.t. \sum_{i=1}^{N} \pi_i = 1)$$

使用拉格朗日乘子法：

$$\mathcal{L}(\pi, \eta) = \sum_{i=1}^{N} \log \pi_i \, P(O, i_1 = q_i \mid \lambda^{(t)}) + \eta \left(\sum_{i=1}^{N} \pi_i - 1 \right)$$

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = \frac{1}{\pi_i} P(O, i_1 = q_i \mid \lambda^{(t)}) + \eta = 0$$

$$\sum_{i=1}^{N} [P(O, i_1 = q_i \mid \lambda^{(t)}) + \pi_i \cdot \eta] = 0$$

$$P(O \mid \lambda^{(t)}) + \eta = 0 \Rightarrow \eta = -P(O \mid \lambda^{(t)})$$

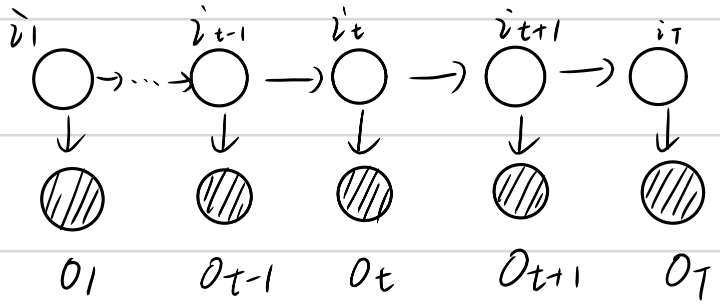
由 $P(O, i_1 = q_i \mid \lambda^{(t)}) + \eta \pi_i = 0$

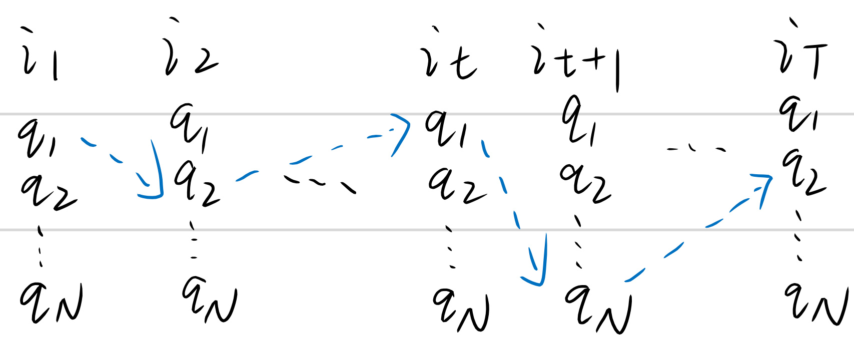
$$\pi_i^{(t+1)} = \frac{1}{\eta} \cdot P(O, i_1 = q_i \mid \lambda^{(t)})$$

$$= \frac{P(O, i_1 = q_i \mid \lambda^{(t)})}{P(O \mid \lambda^{(t)})}$$

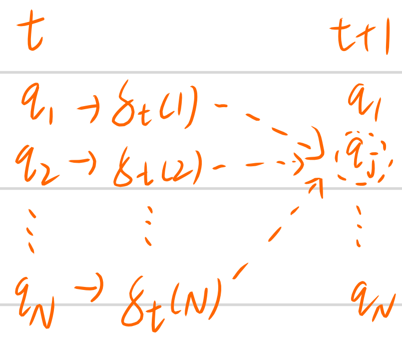
$$\pi^{(t+1)} = (\pi_1^{(t+1)}, \pi_2^{(t+1)}, \cdots \pi_N^{(t+1)})$$

Viterbi 算法





解码问题实际是动态规划问题



$$\delta_t(i) = \max_{i_1, i_2, \cdots i_{t-1}} P(o_1, o_2 \cdots o_t, i_1, i_2, \cdots i_{t-1}, i_t = q_i)$$

$$\delta_{t+1}(j) = \max_{i_1, i_2 \cdots i_t} P(o_1, o_2, \cdots o_t, o_{t+1}, i_1, i_2, \cdots i_t, i_{t+1} = q_j)$$

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

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