白板推导系列(14)　　HMM－背景

Github: CBJerry993

频率派 → 统计机器学习

优化问题 1. model: $f(w)=W^Tx+b$

2. 策略: loss func

3. 算法: GD、SGD、牛顿法、拟牛顿法

贝叶斯派 → 概率图模型

Inference 后验概率 $p(z|x)$ → 积分问题 → MCMC 蒙特卡洛

概率图 { 有向 → Bayesian Network
　　　　无向 → Markov Random Field (Markov Network)

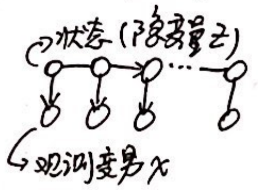
+time (一段话, 一段序列) → Dynamic Model 动态模型 { HMM, Kalman Filter, particle Filter

特点 { x_i之间不是 iid
　　① 横向: time
　　② 纵向: mixture 混合

GMM　$x_1\ x_2 \cdots x_N$　$x_i \overset{iid}{\sim} p(x|\theta)$

z → 离散 $z=1,2,3\cdots k$

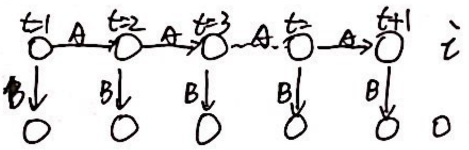
x　$p(x|z)\sim N(\mu,\Sigma)$



+time: 动态

{ 状态离散: HMM
　状态连续: { 线性 → Kalman Filter
　　　　　　非线性 → particle Filter

HMM 结构



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

π: 初始变量

A: 状态转移矩阵

B: 发射矩阵

观测变量 $o, o_1 o_2 \cdots o_t$ → $V=\{v_1 v_2 \cdots v_M\}$ M: 观测值个数

隐状态变量 i $i_1 i_2$ i_t → $Q=\{q_1 q_2 \cdots q_N\}$ N: 离散值

$A=[a_{ij}]$　$a_{ij}=p(i_{t+1}=q_j|i_t=q_i)$ 在

$B=[b_{jk}]$　$b_j(k)=P(O_t=V_k|i_t=q_j)$

π: $i_1=q_1, q_2 \cdots q_n$ 的概率, 记为 $\pi=(p_1, p_2 \cdots p_N)$

两个假设: ① 齐次 Markov 假设

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

② 观察独立假设

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

三个问题

① Evaluation $P(O|\lambda)$ → 已知λ, 计算此概率, 前向/后向算法

② Learning 问题, 如何求 $\lambda=\arg\max P(O|\lambda)$ → EM 算法

③ Decoding 问题, 找到状态序列 $I=i_1 i_2 \cdots i$

使得 $I=\arg\max\limits_{i} P(I|O)$　在什么 I 下, 看到这个序列概率最大

{ 预测 $P(i_{t+1}|O_1 O_2 \cdots O_t)$　下一个隐状态
　滤波 $P(i_t|O_1 O_2 \cdots O_t)$　t 时刻隐状态是什么

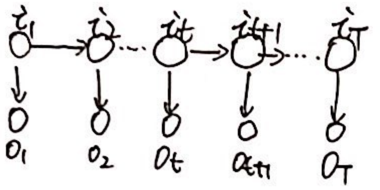
重点是 ③ 问题

{ Kalman Filter
　particle Filter

白板推导系列（14）　HMM — 前向算法

GitHub: CBJerry993

复习HMM



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

π: 初始概率分布 $\pi(\pi_1, \pi_2 \cdots \pi_N)$ $\sum_{i=1}^{N} \pi_i = 1$

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

B: $[b_j(k)]$ 发射矩阵 $b_j(k) = P(o_t = v_k \mid i_t = q_j)$

(2) 两个假设 ① 齐次Markov $P(i_{t+1} \mid i_1 i_2 \cdots i_t, o_1 \cdots o_t) = P(i_{t+1} \mid i_t)$

② 观测独立 $P(o_t \mid i_1, i_2 \cdots i_t, o_1 \cdots o_t) = P(o_t \mid i_t)$

$I = i_1 i_2 \cdots i_T$ 状态序列，$Q = \{q_1 q_2 \cdots q_N\}$ 状态值集合

$O = o_1 o_2 \cdots o_T$ 观测序列，$V = \{v_1 v_2 \cdots v_M\}$ 观测值集合

(3) 三个问题

① Evaluation　Given λ，求 $P(O \mid \lambda) \rightarrow F/B$

② Learning　$\lambda_{MLE} = \arg\max_{\lambda} P(O \mid \lambda) \rightarrow EM$

③ Decoding　$\hat{I} = \arg\max_{i} P(I \mid O, \lambda) \rightarrow$ Viterbi

问题一：

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

$$\because P(I \mid \lambda) = P(i_1 i_2 \cdots i_T \mid \lambda) = P(i_T \mid i_1 i_2 \cdots i_{T-1}, \lambda) \cdot P(i_1 i_2 \cdots i_{T-1} \mid \lambda) = a_{i_{T-1}, i_T} \cdot a_{i_{T-2}, i_{T-1}} \cdots a_{i_1, i_2} \cdot \pi(i_1)$$

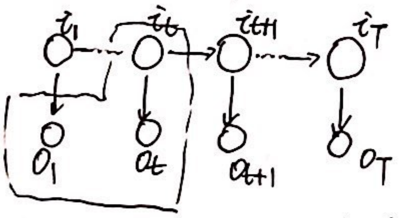
$$= \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1}, i_t}$$

$$P(i_T \mid i_{T-1}) = \cancel{a_{i_{T-1}} i_T} = a_{i_{T-1}, i_T}$$

$$\because P(O \mid I, \lambda) = \prod_{t=1}^{T} b_{i_t}(o_t) = b_{i_1}(o_1) \cdot b_{i_2}(o_2) \cdots b_{i_T}(o_T) \rightarrow$$ 观测独立假设 → 复杂度 $O(N^T)$ 不可解

$$\therefore P(O \mid \lambda) = \underbrace{\sum_I}_{T个求和} \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1}, i_t} \cdot \prod_{t=1}^{T} b_{i_t}(o_t) \overset{展开}{=} \sum_{i_1} \sum_{i_2} \cdots \sum_{i_T} \pi(i_1) \cdot \prod_{t=2}^{T} a_{i_{t-1}, i_t} \prod_{t=1}^{T} b_{i_t}(o_t)$$

Forward Algorithm



记 $\alpha_t(i) = P(o_1 \cdots o_t, i_t = q_i \mid \lambda)$

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

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

$$\alpha_{t+1}(j) = P(o_1 \cdots o_t, o_{t+1}, i_{t+1} = q_j \mid \lambda)$$

$$= \sum_{i=1}^{N} P(o_1 \cdots o_t, o_{t+1}, i_{t+1} = q_j, \underbrace{i_t = q_i}_{加入} \mid \lambda)$$

$$= \sum_{i=1}^{N} P(o_{t+1} \mid o_1 \cdots o_t, i_t = q_i, i_{t+1} = q_j, \lambda) \cdot P(o_1 \cdots o_t, i_t = q_i, i_{t+1} = q_j \mid \lambda)$$

↓假设2 ▲

$$= \sum_{i=1}^{N} P(o_{t+1} \mid i_{t+1} = q_j) \cdot P(i_{t+1} = q_j \mid o_1 \cdots o_t, i_t = q_i, \lambda) \cdot P(o_1 \cdots o_t, i_t = q_i \mid \lambda)$$

↓假设1 ▲

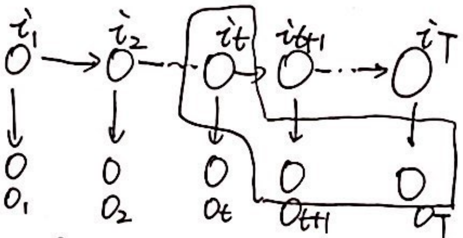
$$= \sum_{i=1}^{N} P(o_{t+1} \mid i_{t+1} = q_j) \cdot P(i_{t+1} = q_j \mid i_t = q_j, \lambda) \cdot \alpha_t(i)$$ 递归式

$$= \sum_{i=1}^{N} b_j(o_{t+1})\, a_{i,j} \cdot \alpha_t(i)$$

白板推导系列(14)　HMM-后向算法

Github: CBJerry993

Evaluation　Given λ, 求 $P(O|\lambda)$



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

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

$$P(O|\lambda) = P(o_1 o_2\cdots o_T|\lambda)$$

$$= \sum_{i=1}^N P(o_1 o_2\cdots o_T, i_1 = q_i) \quad \text{λ不写了(忽略)}$$

$$= \sum_{i=1}^N P(o_1\cdots o_T \mid i_1 = q_i)\cdot \underbrace{P(i_1 = q_i)}_{\pi_i}$$

$$= \sum_{i=1}^N P(o_1 \mid o_2\cdots o_T, i_1 = q_i)\cdot P(o_2\cdots o_T \mid i_1 = q_i)\pi_i$$

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

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

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

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

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

引入 $\overset{a}{O}\rightarrow\overset{b}{O}\rightarrow\overset{c}{O}$ (证明略)

$P(c|a,b) = P(c|b)$：给定a,b求c，其实与a无关，打断

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

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

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

完！

总结:

$$P(O|\lambda) = \sum_I P(O,I|\lambda) = \sum_{i_1}\cdots\sum_{i_T}\pi_{i_1}\prod_{t=2}^T a_{i_{t-1},i_t}\prod_{t=1}^T b_{i_t}(o_t) \rightarrow O(N^T) \text{ 不可解}$$

$\alpha_t(i) = P(o_1\cdots o_t, i_t = q_i|\lambda)$

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

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

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

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

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

机器学习推导系列(14) HMM Baum-Welch (EM) $\boxed{\text{GitHub: CBJerry993}}$

问题二：Learning $\lambda_{mle} = \arg\max_\lambda P(O|\lambda)$

EM		HMM	
x: 观测	$\to$	O	
z: 隐变量	$\to$	I	$\to$ 离散
θ: 参数	$\to$	λ	

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

HMM: $\lambda^{(t+1)} = \arg\max_\lambda \sum_I \log P(O,I|\lambda)\cdot P(I|O,\lambda^{(t)})$ （积分变Σ，$\because$离散）

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

$G = \cancel{P(\Xi)} = \dfrac{P(I,O|\lambda^{(t)})}{\cancel{P(O,\lambda^{(t)})}}$ 无关

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

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

求$\pi^{(t+1)}$，其它的$A^{(t+1)}$，$B^{(t+1)}$方法同，略。

$\pi^{(t+1)} = \arg\max_\pi Q(\lambda,\lambda^{(t)}) = \arg\max_\pi \sum \left[\log\pi_{i_1}\cdot P(O,I|\lambda^{(t)})\right]$ （把与π无关去掉）

$= \arg\max_\pi \sum_{i_1}\cdots\sum_{i_T}\left[\log\pi_{i_1}\cdot P(O,i_1\cdots i_T|\lambda^{(t)})\right]$ 把Σ放入P里面，对P里面有i_1到i_T，对$i_2\sim i_T$求边缘分布，可以去掉Σ.（$i_2\sim i_T$依次求边缘分布把联合分布的$i_2\sim i_T$都积掉）

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

换下标

$= \arg\max_\pi \sum_{i=1}^N\left[\log\pi_i\cdot P(O,i_1=q_i|\lambda^{(t)})\right]$

$(s.t.\ \sum_{i=1}^N(\pi_i)=1)$ 有约束的优化问题

拉格朗日乘子法.

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

$\dfrac{\partial L}{\partial \pi_i} = \dfrac{1}{\pi_i}P(O,i_1=q_i|\lambda^{(t)}) + \eta = 0 \longrightarrow P(O,i_1=q_i|\lambda^{(t)}) + \eta\pi_i = 0$

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

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

$\therefore \eta = -P(O|\lambda^{(t)})$

$\pi_i = -\dfrac{1}{\eta}P(O,i_1=q_i|\lambda^{(t)})$

$\therefore \pi_i^{(t+1)} = \dfrac{P(O,i_1=q_i|\lambda^{(t)})}{P(O|\lambda^{(t)})}$

找到了循环递推关系.

Baum-Welch 实际上是 EM

提出比EM早.

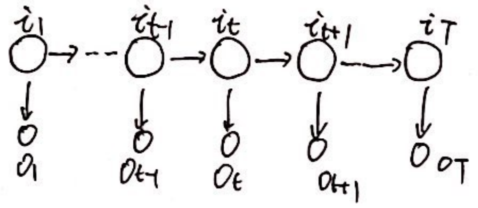
白板推导系列(14)　HMM　Decoding问题 — Viterbi

Github: CBJerry993

问题三: Decoding　$\hat{I} = \arg\max_{I} p(I|O,\lambda)$

预测 $p(I|O)$ 已知t，求t+1. $p(o_{T+1}|o_1\cdots o_T)$ 或 $p(i_{t+1}|o_1\cdots o_t)$

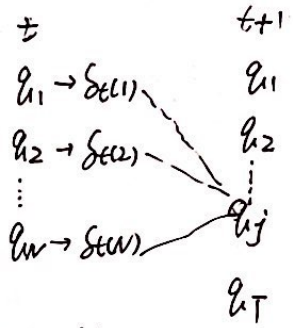
Decoding $p(i_1\cdots i_T|o_1\cdots o_T)$ → 序列，　所以，叫Decoding问题。



每一个时刻，i 都有 $q_N \sim q_1$ 个选择。　N^T

找最优序列，让后验概率最大。　DP

假定t时刻状态是i. 那么到达的最优路径



Viterbi

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

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

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

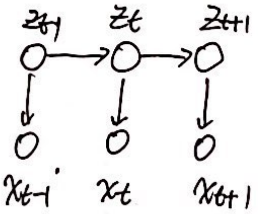
记 上一步是从什么位置到 q_j 的概率

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

白板推导系列(14) HMM — 小结 Filtering, Smoothing, Prediction 问题

$\boxed{\text{Github: CBJerry993}}$

Dynamic model → State space model



$\alpha_t = p(x_1 x_2 \cdots x_t, z_t)$

$\beta_t = p(x_{t+1} x_{t+2} \cdots x_T \mid z_t)$

通识问题

⑤个算法

- Learning (入手点): $\lambda_{mle} = \arg\max_{\lambda} p(x \mid \lambda)$ → EM① → 使用了 F-B 算法 (类似 smoothing)
- Inference (入手点):
  - decoding: $p(z_1 z_2 \cdots z_t \mid x_1 x_2 \cdots x_t)$ → $z = \arg\max_z p(z \mid x)$ → DP → Viterbi②
  - prob of evidence: $p(x \mid \theta) = p(x_1 x_2 \cdots x_T \mid \theta)$ → 如 HMM 的 Evaluation → B/P③ ④
  - filtering: 滤波问题 $p(z_t \mid x_1 x_2 \cdots x_t)$ → online $p(z_1 \mid x_1) \to p(z_2 \mid x_1 x_2) \to p(z_t \mid x_1 x_2 \cdots x_t)$ → F 算法
  - smoothing: 平滑问题 $p(z_t \mid x_1 x_2 \cdots x_T)$ → 拿到所有观测值 → offline 复杂 ⑤ → F-B 算法
  - prediction: 预测问题 $p(z_{t+1} \mid x_1 \cdots x_t)$ → F 算法
    - 或者 $z_{t+1}, z_{t+2} \mid$
    - $p(x_t, x_{t+1} \mid x_1 \cdots x_2 \cdots x_t)$

①

$$\text{Filtering: } p(z_t \mid x_{1:t}) = \frac{p(x_{1:t}, z_t)}{\sum_{z_t} p(x_{1:t}, z_t) \to p(x_{1:t})} \propto p(x_{1:t}, z_t) \to \text{~~ford~~ 前向 F 算法} \to \alpha_t$$

②

$$\text{Smoothing: } p(z_t \mid x_{1:T}) = \frac{p(x_{1:T}, z_t)}{p(x_{1:T}) \to \sum_{z_t} p(x_{1:T}, z_t)} \propto \alpha_t \cdot \beta_t$$

令 $p(x_{1:T}, z_t) = p(x_{1:t}, x_{t+1:T}, z_t) = p(x_{t+1:T} \mid x_{1:t}, z_t) \cdot p(x_{1:t}, z_t)$ → α_t

A C B

概率图模型 $A \to x_{1:t}$　$B \to z_t$　$C \to x_{t+1:T}$，路径 $A \to B \to z_{t+1} \to C$，若 B 断了。

那么 A 与 C 独立，即 $A \perp C \mid B$ → $A \mid B$ (概率，独立) $= p(x_{t+1:T} \mid z_t) \cdot \alpha_t = \beta_t \cdot \alpha_t$

③ prediction

$$p(z_{t+1} \mid x_{1:t}) = \sum_{z_t} p(z_{t+1}, z_t \mid x_{1:t}) = \sum_{z_t} \underbrace{p(z_{t+1} \mid z_t, x_{1:t})}_{\downarrow \text{假设1: } p(z_{t+1} \mid z_t)\text{ 已知}} \cdot \underbrace{p(z_t \mid x_{1:t})}_{\text{filtering}}$$

$$p(x_{t+1} \mid x_{1:t}) = \sum_{z_{t+1}} p(x_{t+1}, z_{t+1} \mid x_{1:t}) = \sum_{z_{t+1}} \underbrace{p(x_{t+1} \mid z_{t+1}, x_{1:t})}_{\downarrow \text{假设2: } p(x_{t+1} \mid z_{t+1})\text{ 已知}} \cdot \underbrace{p(z_{t+1} \mid x_{1:t})}_{\text{filtering}}$$