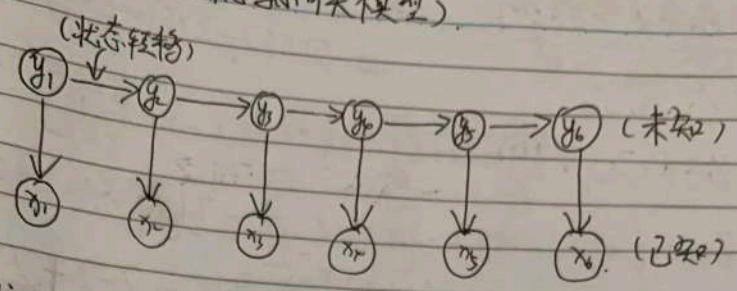
MEMO NO.
DATE

HMM（隐马尔可夫模型）



HMM：

{ ①用于标注问题
  ②是生成模型。

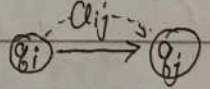
1）定义：

HMM的三要素：

① 状态集合 Q ⟹ 状态转移矩阵 A：

$A=[a_{ij}]_{N\times N}$　N表示状态的数量。

* a_{ij}表示：t时刻处于状态q_i的条件下，在t+1时刻转移到状态q_j的概率。



② 观测集合 V ⟹ 观测概率矩阵 B：

$B=[b_i(k)]_{N\times M}$　M表示观测值的数量

表示：在时刻t处于q_i的条件下，生成观测v_k的概率；即 $b_i(k)=P(O_t=v_k \mid i_t=q_i)$

③ 初始状态概率向量 π：

$\pi=[\pi_i]$

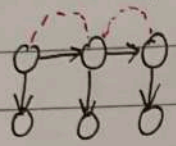
$\pi_i=P(i=q_i)$ 表示在时刻t=1处于q_i的概率。

（只在求解观测状态过程作用）。

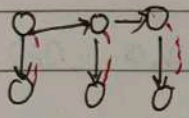
2）两个假设：

①齐次马尔可夫性假设。

任意时刻t的状态只依赖于其前一时刻的状态。



②观测独立性假设。任意时刻的观测，只依赖于该时刻的状态



*需写出以下内容（对HMM）：

①状态的集合。

②观测的集合。

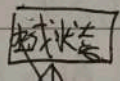
③状态序列和观测序列的长度。

④初始状态概率分布（向量）

⑤状态转移的概率

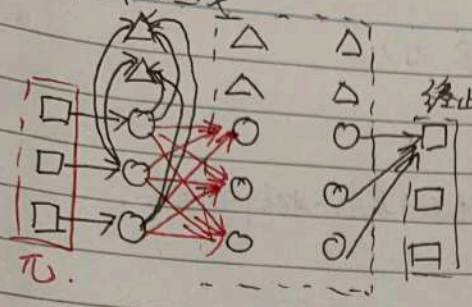
⑥观测概率。

3）观测序列的生成：



MEMO NO.
DATE . .

3) 观测序列的生成



4) HMM的3个基本问题：

1) 概率计算问题：

给定 $\lambda(A, B, \pi)$ 和观测序列 $O=(o_1, o_2, \cdots, o_T)$，计算条件概率 $P(O|\lambda)$

2) 学习问题：

已知观测序列 $O=(o_1, o_2, \cdots, o_T)$ 求 $\lambda(A, B, \pi)$ 的参数.

即：用 MLE 求参数.

3) 预测问题（解码）：

已知模型 $\lambda=(A, B, \pi)$ 和观测序列 $(o_1, o_2, \cdots, o_T)$，求条件概率：$P(I|O, \lambda)$.

5) 概率计算问题，即求 $P(O|\lambda)$.

① 直接计算法：（O：观测序列，I：状态序列）

$P(O, I|\lambda)=P(O|I, \lambda)P(I|\lambda)$ 即：对固定 I，求 $P(O, I|\lambda)$

$P(O|\lambda)=\sum_I P(O, I|\lambda)$ 即：对所以有可能的 I 求和.

时间复杂度：$O(TN^T)$（N表示状态种类，T表示O的次数，此时O很长，故和m（观测值种类）无关）.

MEMO NO.
DATE

② 前向后向算法：(求 $P(O|\lambda)$) (概念22比，只发前与后的)

a) 前向概率：

给定隐马尔可夫模型 λ，定义到时刻 t 部分观测序列为 $o_1, o_2, \cdots, o_t$ 且状态为 q_i 的概率为前向概率，记作：

$$\alpha_t(i) = P(o_1, o_2, \cdots, o_t, i_t = q_i | \lambda).$$

~~表示 q_i，可以~~

i 表示状态，可取所以有的状态值。

b) 算法流程：

输入：隐马尔可夫模型 λ，观测序列 O.

输出：观测序列概率 $P(O|\lambda)$.

(1) 初值 $\alpha_1(i) = \pi_i b_i(o_1)$ $i = 1, 2, \cdots, N$

($t=1$). (N表示状态的种类).

(2) 递推，对 $t = 1, 2, \cdots, T-1$.

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j) a_{ji}\right] b_i(o_{t+1}), \quad i = 1, 2, \cdots, N.$$

观测概率.

(3) 终止：

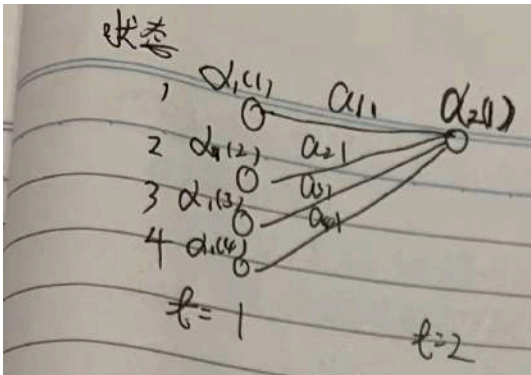
$$P(O|\lambda) = \sum_{i=1}^{N} \alpha_T(i).$$

(T时刻所有状态的概率和).

$\alpha_t(j) a_{ji}$：从 t 时刻，状态 j 转移到状态 i 的概率 (此时，$o_1, \cdots, o_t$ 的概率由 $\alpha_t(j)$ 表示)

$\sum_{j=1}^{N} \alpha_t(j) a_{ji}$ 所有到 $t+1$ 时刻状态为 i 的概率和.
(当前状态，可由之前所有状态转移过来).

$b_i(o_{t+1})$ 在状态 i，观测到 o_{t+1} 的概率.



* 时间复杂度 $O(N^2T)$，
用动态规划一边计算，减少计算量。

③ 后向算法：

(1) 后向概率：

给定隐马尔可夫模型入，定义在时刻 t 状态为 q_i 的条件下，从 t+1 到 T 的部分观测序列为 $o_{t+1}, o_{t+2}, \cdots, o_T$ 的概率为后向概率：

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \cdots, o_T | i_t = q_i, \lambda)$$

(2) 后向算法：

输入：隐马尔可夫模型入，观测序列 O。

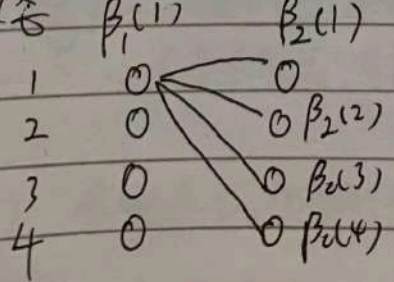
输出：观测序列概率 $P(O|\lambda)$。

(1) $\beta_T(i) = 1 \quad i = 1, 2, \cdots, N$

(2) 对 $t = T-1, T-2, \cdots, 1$

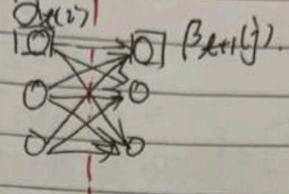
$$\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad i = 1, 2, \cdots, N$$

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



6) HMM观测序列概率 P(O|λ) 统一形式：

$$P(O|\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_t(i)a_{ij}b_j(O_{t+1})\beta_{t+1}(j),\ t=1,2,\cdots,T-1.$$



(某2状态之一个联合概率).

※ 前向后向算法的用处：

给定λ，计算O的概率
(观测值).

7) 学习算法：(求λ(A,B,π)的参数)

监督：MLE

非监督：Baum-Welch算法 (即EM算法).

8) 预测算法：解码 (求 P(I|O,λ)，即状态序列的概率).

近似算法：每个时刻最有可能的状态，只贪心，

维特比算法：全局最优.

9) 关键要点：

①初始时，π表示某一状态的概率值.

②后续求某一时刻t的状态概率由前一时刻t-1到某一状态的概率，再求该状态下对应观测值的概率.

③不同状态之间可能相互转移，即转移矩阵A.

④某一观测值可由不同状态获取，即观测概率B.

⑤先转移再观测(第一个状态由π获得).

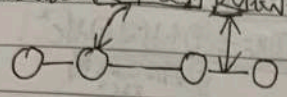
CRF（条件随机场）

1）概率无向图模型：

又称为“马尔可夫随机场”。

① 由图表示概率分布。

② 结点表示随机变量，边表示随机变量之间的依赖关系。



③ 联合概率分布 $P(Y)$ 和表示它的无向图 G 中，随机变量存在以下性质：

成对马尔可夫性
局部马尔可夫性
全局马尔可夫性
} 三者等价。

HMM 的两个假设：

齐次马尔可夫性假设（当前状态仅仅只有一个前面的状态，有向图）

观测独立性假设。

*成对马尔可夫性：

两个变量 Y_u, Y_v，没有边连接，则两个变量在其他结点 O 确定条件下，条件独立：

$$P(Y_u, Y_v | Y_O) = P(Y_u | Y_O) P(Y_v | Y_O)$$

（简单理解为，没有边连接的两个变量，条件独立）。

特点：（设 $P(Y)$ 表示概率无向图的联合概率分布）

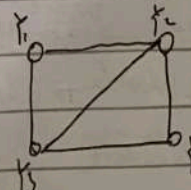
① $P(Y)$ 满足 **马尔可夫性

② 量子因子分解。

2）无向图模型的团与最大团：

① 团：任何两个结点均有边连接的子集

② 最大团：结点数量最多的团。



最大团 $\{Y_1, Y_2, Y_3\}$ 和 $\{Y_2, Y_3, Y_4\}$。

3) 无向图模型的因子分解：

定义：联合概率分布由最大团上的随机变量的正值的乘积形式表示。

表示：(由 Hammersley-Clifford 定理保证)

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C(Y_C)$$

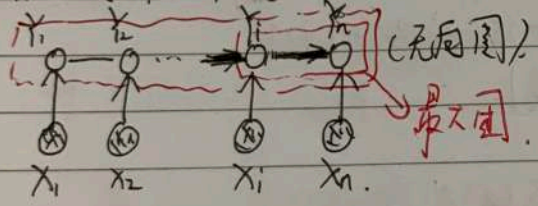
C：为无向图上的最大团。

Y_C：C 对应的随机变量。

Z：规范化因子，$Z=\sum_{Y}\prod_{C}\Psi_C(Y_C)$（所有可能的 Y 的取值和）。

$\Psi_C(Y_C)$：势函数，要求恒正，$\Psi_C(Y_C)=\exp\{-E(Y_C)\}$。

4) 线性链条件随机场。



定义设 $X=(X_1,X_2,\cdots,X_n)$，$Y=(Y_1,Y_2,\cdots,Y_n)$ 均为线性链表示的随机变量序列，若给定 X，且 Y 满足**马尔可夫性**，

即 $P(Y_i|X,Y_1,\cdots,Y_{i-1},Y_{i+1},\cdots Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1})$

$i=1,2,\cdots,n$（在 i 和 n 时只考虑单边）。

相邻节点

则称 $P(Y|X)$ 线性链条件随机场，X：观测序列，Y：状态序列。

MEMO NO.
DATE

5）条件随机场的参数化形式：

设 $P(Y|X)$ 为线性链条件随机场，则在 X 取值为 x 的条件下，随机变量 Y 取值为 y 的条件概率具有如下形式：

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$$

（λ_k：权值，t_k：特征函数；μ_l：权值，s_l：特征函数；$\sum_{i,k}$：位置上的所有特征函数）

（$P(Y)=\frac{1}{Z}\prod_c \Psi_c(Y_c)$）（$Z(x)$的含义参考例11.2）

规范化因子：$Z(x)=\sum_y\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right)$

其中：t_k，s_l 是特征函数

λ_k，μ_l 对应的权值。

* ① $t_k(y_{i-1},y_i,x)$ 依赖当前位置和前一个位置，称为~~状态~~转移特征

② $s_l(y_i,x,i)$ 只是在当前位上，称为状态特征

③ t_k，s_l 一般取 0 或 1.

④ CRF完全依赖于特征函数。

（例子参见 例11.1）

6）CRF的简化形式：

①

$$f_k(y_{i-1},y_i,x,i)=\begin{cases}t_k(y_{i-1},y_i,x,i) & k=1,2,\cdots,K_1\\ s_l(y_i,x,i), & k=K_1+l,\ l=1,2,\cdots,K_2\end{cases}$$

② 对转移特征与状态特征在各个位置求和，记作：

$$f_k(y,x)=\sum_{i=1}^{n}f_k(y_{i-1},y_i,x,i),\ k=1,2,\cdots,K.$$

MEMO NO
DATE . .

③ 权值：

$$w_k=\begin{cases}\lambda_k & k=1,2,\cdots,k_1\\ \mu_l & k=k_1+l;\ l=1,2,\cdots,k_2.\end{cases}$$

④ $P(y|x)=\frac{1}{Z(x)}\exp\sum_{k=1}^{K}w_k f_k(y,x)$

$$Z(x)=\sum_y \exp\sum_{k=1}^{K}w_k f_k(y,x)$$

⑤ 再简化：

$$w=(w_1,w_2,\cdots,w_K)^T$$

$$F(y,x)=(f_1(y,x),f_2(y,x),\cdots,f_K(y,x))^T$$

则 $P_w(y|x)=\frac{\exp(w\otimes F(y,x))}{Z_w(x)}$ ⊗表示内积

$$Z_w(x)=\sum_y \exp(w\otimes F(y,x))$$

7) CRF的矩阵形式：

引入特殊的起点和终点状态标记 $y_0=start$, $y_{n+1}=stop$.

对观测到的x的每一个位置 $i=1,2,\cdots,n+1$, 定义m阶矩阵，

(m标记 y_i 取值的个数，即状态的种类).

$$M_i(x)=[M_i(y_{i-1},y_i|x)]$$

$$M_i(y_{i-1},y_i|x)=\exp(W_i(y_{i-1},y_i|x))$$

$$W_i(y_{i-1},y_i|x)=\sum_{k=1}^{K}w_k f_k(y_{i-1},y_i,x,i).$$

当前位置的特征函数的加权和.

(CRF定义的特征函数)

DATE

于是：

$$\begin{cases} P_w(y|x)=\frac{1}{Z_w(x)}\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x) \\ Z_w(x)=(M_1(x)M_2(x)\cdots M_{n+1}(x))_{start,stop} \end{cases}$$

* 由 $M_i(y_{i-1},y_i|x)=\exp\left(\sum_{k=1}^{K}w_k f_k(y_{i-1},y_i,x,i)\right)$.

则 $\prod_{i=1}^{n+1}M_i(y_{i-1},y_i|x)$

$$=\exp\sum_{i=1}^{n+1}\sum_{k=1}^{K}w_k f_k(y_{i-1},y_i,x,i).$$

$$=\exp\sum_{i,k}w_k f_k(y_{i-1},y_i,x,i).$$

(与式见到 11.2).

8) CRF前后算法.

* CRF, Hmm 前后算法的区别：

当前 CRF：求状态序列的概率.

Hmm：求观测序列的概率

对每个指标 $i=0,1,\cdots,n+1$，定义前向向量 $\alpha_i(x)$：

$$\alpha_0(y|x)=\begin{cases}1, & y=start \\ 0, & 否则.\end{cases}\quad (start起位置).$$

递推公式：

$$\alpha_{i+1}^{T}(y_i|x)=\alpha_i^{T}(y_{i-1}|x)M_i(y_{i-1},y_i|x),\ i=1,2,\cdots,n+1.$$

$(1\times m)\quad (m\times m)$.

即：$\vec{\alpha}_i^{T}(x)=\vec{\alpha}_{i-1}^{T}(x)\vec{M}_i(x)$.

$(1\times m)$.

$\alpha_i(y_i|x)$ 表示在位置 i 的标记是 y_i 并且到位置 i 的前部分标记序列的非规范化概率，y_i 由可取值的 m 个.

MEMO NO.
DATE . .

9) CRF后向算法：

对每个标注位置 $i=0,1,\cdots,n+1$，定义后向向量 $\beta_i(x)$：

$$\beta_{n+1}(y_{n+1}|x)\begin{cases}=1 & y_{n+1}=stop\\ =0 & 否则\end{cases}$$

$$\Rightarrow \beta_i(y_i|x)=M_i(y_i,y_{i+1}|x)\beta_{i+1}(y_{i+1}|x)$$

$$\Rightarrow \vec{\beta}_i(x)=\vec{M}_{i+1}(x)\vec{\beta}_{i+1}(x)$$

$\beta_i(y_i|x)$ 表示在位置 i 的标记为 y_i 并且从 $i+1$ 到 n 的后部分标记序列的非规范化概率。

则

$$Z(x)=\vec{\alpha}_n(x)\cdot\vec{1}=\vec{1}\cdot\vec{\beta}_1(x)$$

$\vec{1}$ 为全1向量。

10) 概率计算：

位置 i 的 y_i 的条件概率：

$$P(Y_i=y_i|x)=\frac{\vec{\alpha}_i^T(y_i|x)\vec{\beta}_i(y_i|x)}{Z(x)}$$

在位置 $i-1$ 与 i 标记 y_{i-1} 和 y_i 的条件概率：

$$P(Y_{i-1}=y_{i-1},Y_i=y_i|x)=\frac{\alpha_{i-1}^T(y_{i-1}|x)M_i(y_{i-1},y_i|x)\beta_i(y_i|x)}{Z(x)}$$

* CRF前后向算法的作用：

① 算 $P(y|x)$ ② 算 $P(y_i|x)$ ③ 算 $P(y_{i-1},y_i|x)$ ④ 还有……

MEMO NO.
DATE

10) CRF预测——维特比算法.

①目的:

给定条件随机场 $P(Y|X)$ 和输入序列(观测)序列 X,

求条件概率最大的输出序列(标记序列) Y^*, 即标注.

②推导:

由之前 $Z_w(x) = \sum_y \exp(w \otimes F(y,x))$

有: $y^* = \arg\max_y P_w(y|x)$ ⊗:内积

$= \arg\max_y \frac{\exp(w \otimes F(y,x))}{Z_w(x)}$

$= \arg\max_y \exp(w \otimes F(y,x))$

$= \arg\max_y [w \otimes F(y,x)]$ (非负单调取对<)

(即对所有位置的过程取和)

③化简:

设 $w = (w_1, w_2, \cdots, w_K)^T$

$F(y,x) = (f_1(y,x), f_2(y,x), \cdots, f_K(y,x))^T$

$f_k(y,x) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i), k = 1, 2, \cdots, K.$

则: $w \otimes F(y,x) = (w_1, w_2, \cdots, w_K)^T \otimes (f_1(y,x), f_2(y,x), \cdots, f_K(y,x))^T$

$= (w_1, w_2, \cdots, w_K)^T \otimes (\sum_{i=1}^{n} f_1(y_{i-1}, y_i, x, i), \cdots, \sum_{i=1}^{n} f_K(y_{i-1}, y_i, x, i))^T$

$= \sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i)$

$= \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, x, i)$

$= \sum_{i=1}^{n} w \cdot F_i(y_{i-1}, y_i, x, i).$

其中 $F_i(y_{i-1}, y_i, x) = (f_1(y_{i-1}, y_i, x, i), f_2(y_{i-1}, y_i, x, i), \cdots, f_K(y_{i-1}, y_i, x, i))^T$

MEMO NO.
DATE . .

④ 步骤：

符号：$\delta_i(j)$ 表示在位置 i 时，label 为 j 的非规范化概率

$\Psi_i(j)$ 表示在位置 i 时，label 为 j 时，上一位置最有可能的 label 是 $\delta_i(j)$ 最大时的上一位置。

(1) 求位置 1 时这 m 种特征的非规范化概率：

$$\delta_1(j) = w F_1(y_0, y_j, x, 1), \quad j = 1, 2, \cdots, m.$$

m 种特征的种类。

(2) 递推求 $\delta_i(l)$（i 表示位置，l 表示 label）

$$\delta_i(l) = \max_{1 \le j \le m} \{\delta_{i-1}(j) + w F_i(y_{i-1}=j, y_i=l, x)\}$$

$$l = 1, 2, \cdots, m.$$

同时记录 $\Psi_i(l) = \arg\max_{1 \le j \le m} \{\delta_{i-1}(j) + w F_i(y_{i-1}=j, y_i=l, x)\}$ $l = 1, 2, \cdots, m.$

(3) 直到位置 n.

$$\max_y (w \cdot f(y, x)) = \max_{1 \le j \le m} \delta_n(j).$$

最终位置为 n 的 label $y_n^* = \arg\max_{1 \le j \le m} \delta_n(j)$

(4) 从终点返回：

$$y_i^* = \Psi_{i+1}(y_{i+1}^*) \quad i = n-1, \cdots, 1.$$

获得最终的最优路径：

$$y^* = (y_1^*, y_2^*, \cdots, y_n^*)^T.$$

(见例 11.3)

MEMO NO.
DATE

1) CRF的学习算法.

{ MLE
正则后的MLE.

* 现在实现算有:
1, 改进的迭代尺度法IIS
2) 梯度下降法
3) 拟牛顿法.

2) NN-CRF.

$$P(y|X)=\frac{e^{s(X,y)}}{\sum_{\tilde{y}\in Y} e^{s(X,\tilde{y})}}$$

$$s(X,y)=\sum_{i=0}^{n} A_{y_i,y_{i+1}}+\sum_{i=1}^{n} P_{i,y_i}$$

LSTM hidden state.

i: 位置
y_i: label
A: 转移矩阵.
$A_{i,j}$: label i 转移到 label j 的分数.