人工智能的数学基础

华东师范大学 数学科学学院 黎芳(教授) 2019年11月18日

Chapter 11 条件随机场（conditional random field，CRF）

Table of Contents

11.1 概率无向图模型

概率无向图模型又称为马尔可夫随机场，是一个可以由无向图表示的联合概率分布。

11.1.1 模型定义

图 (graph) 是由结点 (node) 及连接结点的边 (edge) 组成的集合．结点和边分别记作v和e ，结点和边的集合分别记作V和E, 图记作$G=(V,E)$．无向图是指边没有方向的图。

概率图模型 (probabilistic graphical model) 是由图表示的概率分布． 设有联合概率分布$P(Y)$，$Y\in\mathcal{Y}$是一组随机变量． 由无向图$G=(V,E)$表示概率分布 $P(Y)$， 即在图G中， 结点$v\in V$表示一个随机变量$Y_v, Y=(Y_v)_{v\in V}$. 边$e\in E$表示随机变量之间的概率依赖关系．

成对马尔可夫性； 设u和v是无向图 G 中任意两个没有边连接的结点， 结点u和v分别对应随机变量Y_u和Y_v, 其他所有结点为O， 对应的随机变量组是Y_O．成对马尔可夫性是指给定随机变量组Y_O的条件下随机变量Y_u和Y_v是条件独立的， 即

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

局部马尔可夫性：设$v\in V$是无向团G中任意一个结点，W是与v有边连接的所有结点，O是v和W以外的其他所有结点．v表示的随机变量是Y_v, W表示的随机变量组是Y_W, O表示的随机变量组是Y_O. 局部马尔可夫是指在给定随机变量组Y_W的条件下随机变量Y_v与随机变量组Y_O是独立的，即

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

在$P(Y_O|Y_W) > 0$时，等价地，$P(Y_v|Y_w) = P(Y_v|Y_w, Y_O)$

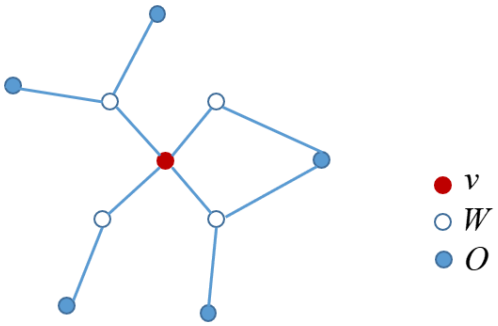

图11.1 局部马尔可夫性

全局马尔可夫性

$$P(Y_A, Y_B|Y_C) = P(Y_A|Y_C)P(Y_B|Y_C)$$

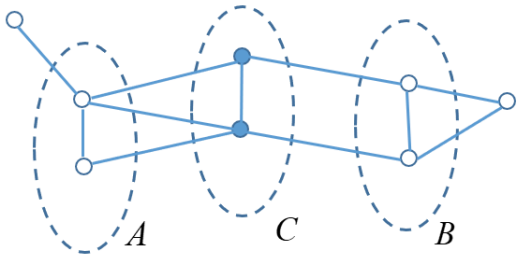

图11.2 全局马尔可夫性

上述成对的、局部的、全局的马尔可夫性定义是等价的。

定义11.1（概率无向图模型）设有联合概率分布$P(Y)$，由无向图$G = (V, E)$表示，在图G中，结点表示随机变量，边表示随机变量之间的依赖关系。如果联合概率分布$P(Y)$满足成对、局部或全局马尔可夫性，就称此联合概率分布为概率无向图模型(probabilityundirected graphical model)，或马尔可夫随机场(Markov random field)。

11.1.2 概率无向图模型的因子分解

定义11.2（团与最大团）无向图G中任何两个结点均有边连接的结点子集称为团(clique). 若C是无向图G的一个团，并且不能再加进任何一个G的结点使其成为一个更大的团，则称此C为最大团(maximal clique).

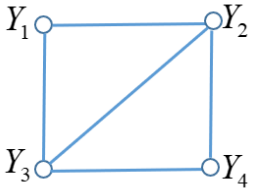

图11.3 无向图的团和最大团

概率无向图模型的因子分解：将概率无向图模型的联合概率分布表示为其最大团上的随机变量的函数的乘积形式的操作。

定义势函数：$\Psi_C(Y_C) = \exp\{-E(Y_C)\}$

定理11.1（Hammersley-Clifford定理）概率无向图模型的联合概率分布P(Y)可以表示为如下形式：

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

$$Z = \sum_Y \prod_C \Psi_C(Y_C)$$

其中，C是无向图的最大团，Y是C的结点对应的随机变量，$\Psi_C(Y_C)$是C上定义的严格正函数，乘积是在无向图所有的最大团上进行的.

11.2 条件随机场的定义与形式

11.2.1 条件随机场的定义

首先定义一般的条件随机场，然后定义线性链条件随机场。

定义11.3（条件随机场）设X与Y是随机变量，$P(Y|X)$是在给定X的条件下Y的条件概率分布。若随机变量Y构成一个由无向图场$G = (V, E)$表示的马尔可夫随机场，即

$$P(Y_v|X, Y_w, w \neq v) = P(Y_v|X, Y_w, w \sim v)$$

对任意结点v成立，则称条件概率分布$P(Y|X)$为条件随机场. $w \sim v$表示在图$G = (V, E)$中与v有边连接的所有结点w.

$$G = (V = \{1, 2, \cdots, n\}, \quad E = \{(i, i+1)\}), \quad i = 1, 2, \cdots, n-1$$

$$X = (X_1, X_2, \cdots, X_n), \quad Y = (Y_1, Y_2, \cdots, Y_n)$$

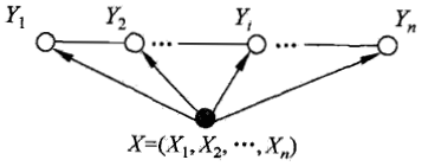

图 11.4 线性链条件随机场

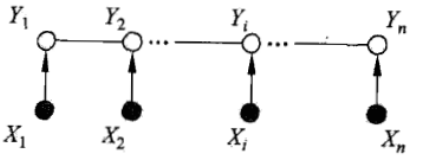

图 11.5 X 和 Y 有相同的图结构的线性链条件随机场

定义11.4（线性链条件随机场）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，若在给定随机变量序列X的条件下，随机变量序列Y的条件概率分布$P(Y|X)$构成条件随机场，即满足马尔可夫性

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

则称$P(Y|X)$为线性链条件随机场．在标注问题中，X表示输入观测序列，Y表示对应的输出标记序列或状态序列．

11.2.2 条件随机场的参数化形式

定理11.2（线性链条件随机场的参数化形式）设$P(Y|X)$为线性链条件随机场，则在随机变量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：

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)\right) \quad (11.10)$$

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

t_k,s_l是特征函数，λ_k,μ_l是对应的权值．$Z(x)$是归一化因子。

t_k是定义在边上的特征，s_l是定义在顶点上的特征。

例**11.1** 设有一标注问题：输入观测序列为$X=(X_1,X_2,X_3)$，输出标记序列为$Y=(Y_1,Y_2,Y_3)$，Y_1,Y_2,Y_3取值于$\mathcal{Y}=\{1,2\}$. 定义特征和权值如下

$$t_1=t_1(y_{i-1}=1,y_i=2,x,i),i=2,3 \quad \lambda_1=1$$

$$t_1(y_{i-1},y_i,x,i)=\begin{cases}1, & y_{i-1}=1,y_i=2,x,i,(i=2,3)\\ 0, & \text{其他}\end{cases}$$

$$\begin{aligned}
&t_2=t_2(y_1=1,y_2=1,x,2) && \lambda_2=0.5\\
&t_3=t_3(y_2=2,y_3=1,x,3) && \lambda_3=1\\
&t_4=t_4(y_1=2,y_2=1,x,2) && \lambda_4=1\\
&t_5=t_5(y_2=2,y_3=2,x,3), && \lambda_5=0.2\\
&s_1=s_1(y_1=1,x,1), && \mu_1=1\\
&s_2=s_2(y_i=2,x,i),i=1,2 && \mu_2=0.5\\
&s_3=s_3(y_i=1,x,i),i=2,3 && \mu_3=0.8\\
&s_4=s_4(y_3=2,x,3), && \mu_4=0.5
\end{aligned}$$

对给定的观测序列x，求标记序列$y=\{1,2,2\}$的非规范化条件概率。

解：由式(11.10), 线性链条件随机场模型为

$$P(y|x)\propto\exp\left[\sum_{k=1}^{5}\lambda_k\sum_{i=2}^{3}t_k(y_{i-1},y_i,x,i)+\sum_{k=1}^{4}\mu_k\sum_{i=1}^{3}s_k(y_i,x,i)\right]$$

对给定的观测序列x，标记序列$y=\{1,2,2\}$的非规范化条件概率为 $P(y_1=1,y_2=2,y_3=2|x)\propto\exp(3.2)$

11.2.3 条件随机场的简化形式

为简便起见，首先将转移特征和状态特征及其权值用统一的符号表示．设有 K_1个转移特征，K_2个状态特征，$K=K_1+K_2$，记

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

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

用w_k表示特征$f_k(y, x)$的权值，即

$$w_k = \begin{cases} \lambda_k, & k = 1, 2, \cdots, K_1 \\ \mu_i, & k = K_1 + l; l = 1, 2, \cdots, K_2 \end{cases}$$

条件随机场可以表示为

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

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

若以$w = (w_1, w_2, \cdots, w_K)^{\mathrm{T}}$表示权值向量，以$F(x, y)$表示全局特征向量，即

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

则条件随机场可以写成向量内积形式

$$P_w(y|x) = \frac{\exp(w \cdot F(y, x))}{Z_w(x)}$$

$$Z_w(x) = \sum_{y} \exp(w \cdot F(y, x))$$

11.2.4 条件随机场的矩阵形式

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

对观测序列x的每一个位置$i = 1, 2, \cdots, n+1$, 由于y_{i-1}和y_i在m个标记中取值，可以定义一个m阶矩阵随机变量

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

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

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

这样，给定观测序列 x, 相应标记序列 y 的非规范化概率可以通过 n+l 个矩阵的适当元素的乘积 $\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|x)$ 表示，于是，条件概率$P_w(y|x)$是

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

$$Z_w(x) = [M_1(x)M_2(x)\cdots M_{n+1}(x)]_{\text{start,stop}} \quad (11.25)$$

规范化因子$Z_w(x)$是以start为起点，以stop为终点通过状态的所有路径$y_1y_2\cdots y_n$的非规范化概率$\prod_{i=1}^{n+1} M_i(y_{i-1}, y_i|x)$之和.

例**11.2** 给定一个由图11.6 所示的线性链条件随机场， 观测序列x，状态序列$y, i = 1, 2, 3, n = 3$，标记$y_i \in \{1, 2\}$，假设$y_0 = \text{start} = 1, y_4 = \text{stop} = 1$.各个位置的随机矩阵$M_1(x), M_2(x), M_3(x), M_4(x)$分别是

$$M_1(x) = \begin{bmatrix} a_{01} & a_{02} \\ 0 & 0 \end{bmatrix}, \quad M_2(x) = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

$$M_3(x) = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}, \quad M_4(x) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}$$

试求状态序列y以start为起点stop为终点所有路径的非规范化概率及规范化因子.

解：首先计算图11.6中从start到stop 对应于$y = (1, 1, 1), y = (1, 1, 2), \cdots, y = (2, 2, 2)$各路径的非规范化概率分别是

$$a_{01}b_{11}c_{11}, \quad a_{01}b_{11}c_{12}, \quad a_{01}b_{12}c_{21}, \quad a_{01}b_{12}c_{22}$$
$$a_{02}b_{21}c_{11}, \quad a_{02}b_{21}c_{12}, \quad a_{02}b_{22}c_{21}, \quad a_{02}b_{22}c_{22}$$

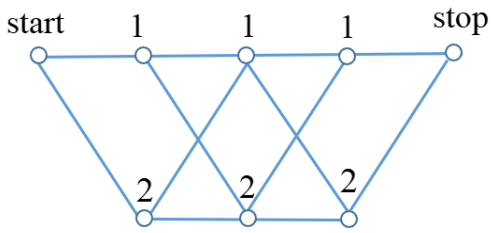

图11.6 状态路径

然后按式(11.25)求规范化因子。通过计算矩阵乘积$M_1(x), M_2(x), M_3(x), M_4(x)$可知，其第1行第1列的元素为

$$a_{01}b_{11}c_{11} + a_{02}b_{21}c_{11} + a_{01}b_{12}c_{21} + a_{02}b_{22}c_{22}$$
$$+a_{01}b_{11}c_{12} + a_{02}b_{21}c_{12} + a_{01}b_{12}c_{22} + a_{02}b_{22}c_{21}$$

恰好等于从start到stop的所有路径的非规范化概率之和， 即规范化因子$Z(x)$.

11.3 条件随机场的概率计算问题

条件随机场的概率计算问题是给定条件随机场$P(Y|X)$， 输入序列x和输出序列y，计算条件概率$P(Y_i = y_i|x), P(Y_{i-1} = y_{i-1}, Y_i = y_i|x)$以及相应的数学期望的问题。

11.3.1 前向-后向算法

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

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

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

$$\alpha_i^{\mathrm{T}}(x)=\alpha_{i-1}^{\mathrm{T}}(x)M_i(x)$$

同样，定义后向向量$\beta_i(x)$:

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

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

$$\beta_i(x)=M_{i+1}(x)\beta_{i+1}(x)$$

11.3.2 概率计算

$$P(Y_i=y_i|x)=\frac{\alpha_i^{\mathrm{T}}(y_i|x)\beta_i(y_i|x)}{Z(x)}$$

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

$$Z(x)=\alpha_n^{\mathrm{T}}(x)\mathbf{1}=\mathbf{1}\beta_1(x)$$

11.3.3 期望值计算

$$\begin{aligned}E_{P(Y|X)}[f_k]&=\sum_y P(y|x)f_k(y,x)\\&=\sum_{i=1}^{n+1}\sum_{y_{i-1}y_i}f_k(y_{i-1},y_i,x,i)\frac{\alpha_{i-1}^{\mathrm{T}}(y_{i-1}|x)M_i(y_{i-1},y_i|x)\beta_i(y_i|x)}{Z(x)}\\&k=1,2,\cdots,K\end{aligned}$$

$$Z(x)=\alpha_n^{\mathrm{T}}(x)\cdot\mathbf{1}$$

假设经验分布为$\widetilde{P}(X)$，特征函数f_k关于联合分布$P(X,Y)$的期望是

$$
\begin{aligned}
E_{P(X,Y)}[f_k] &= \sum_{x,y} P(x,y) \sum_{i=1}^{n+1} f_k(y_{i-1}, y_i, x, i) \\
&= \sum_{x} \widetilde{P}(x) \sum_{y} P(y|x) \sum_{i=1}^{n+1} f_k(y_{i-1}, y_i, x, i) \\
&= \sum_{x} \widetilde{P}(x) \sum_{i=1}^{n+1} \sum_{y_{i-1} y_i} f_k(y_{i-1}, y_i, x, i) \frac{\alpha_{i-1}^{\mathrm{T}}(y_{i-1}|x) M_i(y_{i-1}, y_i|x) \beta_i(y_i|x)}{Z(x)} \\
& k = 1, 2, \cdots, K
\end{aligned}
$$

11.4 条件随机场的学习算法

11.4.1 改进的迭代尺度法

训练数据的对数似然函数为

$$
L(w) = L_{\widetilde{p}}(P_{\mathrm{w}}) = \log \prod_{x,y} P_{\mathrm{w}}(y|x)^{\widetilde{P}(x,y)} = \sum_{x,y} \widetilde{P}(x,y) \log P_{\mathrm{w}}(y|x)
$$

$$
\begin{aligned}
L(w) &= \sum_{x,y} \widetilde{P}(x,y) \log P_w(y|x) \\
&= \sum_{x,y} \left[\widetilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \widetilde{P}(x,y) \log Z_w(x) \right] \\
&= \sum_{x,y} \widetilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \sum_{x} \widetilde{P}(x) \log Z_w(x) \\
&= \sum_{j=1}^{N} \widetilde{P}(x_j, y_j) \sum_{k=1}^{K} w_k f_k(y_j, x_j) - \sum_{j=1}^{N} \widetilde{P}(x_j) \log Z_w(x_j)
\end{aligned}
$$

改进的迭代尺度法通过迭代的方法不断优化对数似然函数改变量的下界， 达到极大化对数似然函数的目的.假设模型的当前参数向量为$w = (w_1, w_2, \cdots, w_K)^T$,向量的增量为$\delta = (\delta_1, \delta_2, \cdots, \delta_K)^{\mathrm{T}}$,更新参数向置为$w + \delta$. 在每步迭代过程中，改进的迭代尺度法通过依次求解式(11.36)和式(11.37)，得到δ， 推导可参考6.3.l节.

关于转移特征t_k的更新方程为

$$
\begin{aligned}
E_{\widetilde{P}[t_k]} &= \sum_{x,y} \widetilde{P}(x,y) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \\
&= \sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \exp\left(\delta_k T(x,y)\right) \\
& k = 1, 2, \cdots; K_1
\end{aligned}
\tag{11.36}
$$

关于状态特征s_l的更新方程为

$$
\begin{aligned}
E_{\widetilde{P}}[s_l] &= \sum_{x,y} \widetilde{P}(x,y) \sum_{i=1}^{n+1} s_l(y_i, x, i) \\
&= \sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n} s_l(y_i, x, i) \exp\left(\delta_{K_1+l} T(x,y)\right) \\
& l = 1, 2, \cdots, K_2
\end{aligned}
\tag{11.37}
$$

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

算法11.1（条件随机场模型学习的改进的迭代尺度法）

输入：特征函数$t_1, t_2, \cdots, t_{K_1}$，　$s_1, s_2, \cdots, s_{K_2}$；经验分布$\widetilde{P}(x,y)$；

输出：参数估计值$\hat{w}$，模型$P_{\hat{w}}$

(1)对所有$k \in \{1, 2, \cdots, K\}$，取初值$w_k = 0$

(2)对每一$k \in \{1, 2, \cdots, K\}$：

(a)当$k \in \{1, 2, \cdots, K\}$时，令$\delta_k$是方程

$$
\sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \exp\left(\delta_k T(x,y)\right) = E_{\widetilde{P}}[t_k]
$$

的解：当$k = K_1 + l$，　$l = 1, 2, \cdots, K_2$时，令δ_{K_1+l}是方程

$$
\sum_{\mathbf{x},y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n} s_l(y_i, x, i) \exp\left(\delta_{K_1+l} T(x,y)\right) = E_{\widetilde{P}}[s_l]
$$

的解：

(b)更新w_k值：$w_k \leftarrow w_k + \delta_k$

(3)如果不是所有w_k都收敛，重复步骤(2).

定义$s(x,y) = S - \sum_{i=1}^{n+1} \sum_{k=1}^{K} f_k(y_{i-1}, y_i, x, i)$，S足够大，使$s \geq 0$.特征总数取为S.

对于转移特征t_k，δ_k的更新方程是

$$
\sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \exp\left(\delta_k S\right) = E_{\widetilde{P}}[t_k]
$$

$\delta_k = \frac{1}{S} \log \frac{E_{\widetilde{P}}[t_k]}{E_P[t_k]}$，　其中

$$E_P(t_k) = \sum_x \widetilde{P}(x) \sum_{i=1}^{n+1} \sum_{i=1}^{n+1} \sum_{x=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \frac{\alpha_{i-1}^{\mathrm{T}}(y_{k-1}|x) M_i(y_{i-1}, y_i|x) \beta_i(y_i|x)}{Z(x)}$$

同样由式(11.37)，对于状态特征s_l，δ_k的更新方程是

$$\sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n} s_i(y_i, x, i) \exp\left(\delta_{K_j+l} S\right) = E_{\widetilde{P}}[s_l]$$

$$\delta_{K_1+l} = \frac{1}{S} \log \frac{E_{\ddot{P}}[s_l]}{E_P[s_l]}$$

$$E_P(s_l) = \sum_x \widetilde{P}(x) \sum_{i=1}^{n} \sum_{y_i} s_l(y_i, x, i) \frac{\alpha_i^{\mathrm{T}}(y_i|x) \beta_i(y_i|x)}{Z(x)}$$

以上算法称为算法S. 在算法S中需要使常数S取足够大，这样一来，每步迭代的增量向量会变大，算法收敛会变慢．算法T试图解决这个问题．算法T对每个观测序列x计算其特征总数最大值T(x),

$$T(x) = \max_y T(x, y)$$

利用前向－后向递推公式，可以很容易地计算$T(x) = t$. 这时，关于转移特征参数的更新方程可以写成：

$$\begin{aligned}
E_{\widetilde{P}}[t_k] &= \sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \exp\left(\delta_k T(x)\right) \\
&= \sum_x \widetilde{P}(x) \sum_y P(y|x) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) \exp\left(\delta_k T(x)\right) \\
&= \sum_{i=1}^{n} \widetilde{P}(x) a_{k,t} \exp\left(\delta_k t\right) \\
&= \sum_{t=0}^{T_{max}} a_{k,i} \beta_k^t
\end{aligned}$$

$$\begin{aligned}
E_{\widetilde{P}}[s_l] &= \sum_{x,y} \widetilde{P}(x) P(y|x) \sum_{i=1}^{n} s_l(y_i, x, i) \exp\left(\delta_{\kappa_1+l} T(x)\right) \\
&= \sum_x \widetilde{P}(x) \sum_y P(y|x) \sum_{i=1}^{n} s_l(y_i, x, i) \exp\left(\delta_{k_1+l} T(x)\right) \\
&= \sum_x \widetilde{P}(x) b_{l,t} \exp\left(\delta_k t\right) \\
&= \sum_{t=0}^{T_{max}} b_{l,t} \gamma_l
\end{aligned}$$

11.4.2 拟牛顿法

对于条件随机场模型

$$P_w(y|x)=\frac{\exp\left(\sum_{i=1}^{n}w_if_i(x,y)\right)}{\sum_y\exp\left(\sum_{i=1}^{n}w_if_i(x,y)\right)}$$

学习的优化目标函数是

$$\min_{w\in\mathbb{R}^n}f(w)=\sum_x\widetilde{P}(x)\log\sum_y\exp\left(\sum_{i=1}^{n}w_if_i(x,y)\right)-\sum_{x,y}\widetilde{P}(x,y)\sum_{i=1}^{n}w_if_i(x,y)$$

其梯度函数是

$$g(w)=\sum_{x,y}\widetilde{P}(x)P_w(y|x)f(x,y)-E_{\widetilde{P}}(f)$$

算法11.2（条件随机场模型学习的**BFGS**算法）

输入：特征函数$f_1,f_2,\cdots,f_n$；经验分布$\widetilde{P}(X,Y)$.

输出：最优参数值$\widehat{w}$，最优模型$P_{\hat{w}}(y[x)$.

- (1) 选定初始点$w^{(0)}$，取B_0为正定对称矩阵，置$k=0$
- (2) 计算$g_k=g(w^{(k)})$. 若$g_k=0$，则停止计算；否则转(3)
- (3) 由$B_kp_k=-g_k$，求出p_k
- (4) 一维搜索：求λ_k使得 $f\left(w^{(k)}+\lambda_kp_k\right)=\min_{\lambda>0}f\left(w^{(k)}+\lambda p_k\right)$
- (5) $w^{(k+1)}=w^{(k)}+\lambda_kp_k$
- (6) 计算 $g_{k+1}=g(w^{(k+1)})$，若$g_k=0$，则停止计算；否则按下式求出B_{k+1}：
- $$B_{k+1}=B_k+\frac{y_ky_k^T}{y_k^{\mathrm{T}}}\delta_k^T-\frac{B_k\delta_k\delta_k^{\mathrm{T}}B_k}{\delta_k^{\mathrm{T}}B_k\delta_k},y_k=g_{k+1}-g_k,\quad\delta_k=w^{(k+1)}-w^{(k)}.$$
- (7) 置$k=k+1$，转(3).

11.5 条件随机场的预测算法

$$\begin{aligned}y^*&=\arg\max_y P_w(y|x)\\&=\arg\max_y\frac{\exp(w\cdot F(y,x))}{Z_w(x)}\\&=\arg\max_y\exp(w\cdot F(y,x))\\&=\arg\max_y(w\cdot F(y,x))\end{aligned}$$

于是，条件随机场的预测问题成为求非规范化概率最大的最优路径问题

$$\max_y(w\cdot F(y,x))$$

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

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

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

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

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

算法11.3（条件随机场预测的维特比算法）

输入．模型特征向量$F(y,x)$和权值向量w，观测序列$x=(x_1,x_2,\ldots,x_n)$；

输出：最优路径$y^*=\left(y_i^*,y_2^*,\cdots,y_n^*\right)$.

(1) 初始化

$$\delta_1(j)=w\cdot F_1(y_0=\text{start},y_1=j,x),\quad j=1,2,\cdots,m$$

(2) 递推：对$i=2,3,\cdots,n$

$$\delta_i(l)=\max_{1\le j\le m}\{\delta_{i-1}(j)+w\cdot F_i(y_{i-1}=j,y_i=l,x)\},\quad l=1,2,\cdots,m$$

$$\Psi_i^{(l)}=\arg\max_{l\le j\le m}\{\delta_{t-1}(j)+w\cdot F_i(y_{i-l}=j,y_l=l,x)\},\quad l=1,2,\cdots,m$$

(3) 终止

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

$$y_n^*=\arg\max_{1\le j\le m}\delta_n(j)$$

(4) 返回路径

$$y_i^*=\Psi_{i+1}\left(y_{i+1}^*\right),\quad i=n-1,n-2,\cdots,1$$

求得最优路径$y^*=\left(y_1^*,y_2^*,\cdots,y_n^*\right)$.

例**11.3** 在例11.1中，用维特比算法求给定的输入序列（观测序列）x对应的最优输出序列（标记序列）$y^*=\left(y_1^*,y_2^*,y_3^*\right)$.

解：利用维特比算法求解最优路径问题

$$\max\sum_{i=1}^{3}w\cdot F_i(y_{i-1},y_i,x)$$

(1) 初始化

$$\delta_1(j) = w \cdot F_1(y_0 = \text{ start }, y_1 = j, x), \quad j = 1, 2$$

$$i = 1, \quad \delta_1(1) = 1, \quad \delta_2(2) = 0.5$$

(2) 递推:

$$i = 2 \quad \delta_2(l) = \max_j \{\delta_1(j) + w \cdot F_2(j, l, x)\}$$

$$\delta_2(1) = \max\{1 + \lambda_2 t_2 + \mu_3 s_3, 0.5 + \lambda_4 t_4 + m u_3 s_3\} = 2.4, \quad \Psi_2(1) = 1$$

$$\delta_2(2) = \max\{1 + \lambda_1 t_1 + \mu_2 s_2, 0.5 + \mu_2 s_2\} = 2.5, \quad \Psi_2(2) = 1$$

$$i = 3 \quad \delta_3(l) = \max_j \{\delta_2(j) + w \cdot F_3(j, l, x)\}$$

$$\delta_3(1) = \max\{2.4 + \mu_5 s_5, 2.5 + \lambda_3 t_3 + \mu_3 s_3\} = 4.3, \quad \Psi_3(1) = 2$$

$$\delta_3(2) = \max\{2.4 + \lambda_1 t_1 + \mu_4 s_4, 2.5 + \lambda_5 t_5 + \mu_4 s_4\} = 3.9, \quad \Psi_3(2) = 1$$

(3) 终止

$$\max_y (w \cdot F(y, x)) = \max \delta_3(l) = \delta_3(1) = 4.3$$

$$y_3^* = \arg\max_l \delta_3(l) = 1$$

(4) 返回路径

$$y_2^* = \Psi_3\left(y_3^*\right) = \Psi_3(1) = 2$$

$$y_1^* = \Psi_2\left(y_2^*\right) = \Psi_2(2) = 1$$

求得最优路径 $y^* = \left(y_1^*, y_2^*, y_3^*\right) = (1, 2, 1)$.

作业

习题11.2