条件随机场(CRF)简介

条件随机场(conditional random field, CRF)是给定一组输入随机变量条件下另一组输出随机变量的条件概率分布模型，其特点是假设输出随机变量构成马尔科夫随机场。

概率无向图模型

概率无向图模型，又称马尔科夫随机场，是一个可以由无向图表示的联合概率分布。

模型定义

图是由结点及边组成的集合，记作v, e，其集合为V, E， 图$G=(V,E)$。

概率图模型是由图表示的概率分布，设有联合概率分布$P(Y)$，Y是一组随机变量。由无向图G表示概率分布$P(Y)$，即在图G中，结点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。

给定一个联合概率分布$P(Y)$和表示它的无向图G，首先定义无向图表示的随机变量之间存在的成对马尔科夫性(pairwise Markov property)、局部马尔科夫性(local Markov property)和全局马尔科夫性(global Markov property)。

成对马尔科夫性：设u, v是任意两个没有边连接的结点，分别对应随机变量Y_u, Y_v，其他所有结点为O，对应的随机变量组是Y_O。成对马尔科夫性是指给定随机变量组Y_O的条件下随机变量Y_u, Y_v是条件独立的，即

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

局部马尔科夫性：设v是无向图G中任意一个结点，W是与v有边相连的所有结点，O是v, W以外的所有结点。局部马尔科夫性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

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

全局马尔科夫性：设结点集合A, B是在无向图中被结点集合C分开的任意结点集合。全局马尔科夫性指给定随机变量组Y_C条件下随机变量组Y_A, Y_B是条件独立的，即

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

上述成对的、局部的、全局的马尔科夫性定义是等价的。

定义(概率无向图模型)：设有联合概率分布$P(Y)$，由无向图G表示，在图中，结点表示随机变量，边表示之间的依赖关系。如果联合概率分布$P(Y)$满足成对、局部或全局马尔科夫性，就称此联合概率分布为概率无向图模型，或马尔科夫随机场。

实际上，我们更关心的是如何求其联合概率分布。对给定的概率无向图模型，我们希望将整体的联合概率写成若干子联合概率的乘积的形式，也就是将联合概率进行因子分解。

概率无向图模型的因子分解

定义(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是一个团，并且不能再加进任何一个G的结点成为一个更大的团，则C为最大团。

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，称为概率无向图模型的因子分解。

给定概率无向图模型，设其无向图为G，C为最大团，Y_C表示其对应的随机变量。那么概率无向图的联合概率分布$P(Y)$可写作图中所有最大团C上的函数$\Psi_C(Y_C)$的乘积形式，即

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

函数$\Psi_C(Y_C)$称为势函数，这里要求势函数$\Psi_C(Y_C)$是严格正的，通常定义为指数函数：

$$\Psi_C(Y_C)=\exp(-E(Y_C))$$

定理(Hammersley-Clifford)：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_C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。

条件随机场的定义与形式

条件随机场的定义

条件随机场是给定随机变量X条件下，随机变量Y的马尔科夫随机场。这个介绍定义在线性链上的特殊条件随机场，称为线性链条件随机场。线性链条件随机场可以用于标注等问题。这时，在条件概率模型$P(Y|X)$中，Y是输出变量，X是输入变量。学习时，利用训练数据集通过极大似然估计或正则化的极大似然估计得到条件概率模型$\hat{P}(Y|X)$；预测时，对于给定的输入序列x，求出条件概率$\hat{P}(y|x)$最大的输出序列$\hat{y}$。

定义(条件随机场)：设X,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G表示的马尔科夫随机场，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中与结点v有边连接的所有结点w。$w\neq v$表示结点v以外的所有结点。

这里没有要求X,Y有相同的结构。现实中，一般假设由相同的图结构，即

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

此时，最大团是相邻两个结点的集合。

定义(线性链条件随机场)：设$X=\{X_1,X_2,\cdots,X_n\},Y=\{Y_1,\cdots,Y_n\}$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，即满足马尔科夫性：

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

则称$P(Y|X)$为线性链条件随机场。在标注问题中，X为观测序列，Y为对应的标记序列或状态序列。

条件随机场的参数化形式

定理(线性链条件随机场的参数化形式)：设$P(X|Y)$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$Y=y$的条件概率具有如下形式：

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

其中，

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

式中，t_k,s_l是特征函数，λ_k,μ_l为对应的权重，$Z(x)$是规范化因子，求和是在所有可能的输出序列上进行的。

上式是线性链条件随机场模型的基本形式，表示给定输入序列x，对输出序列y预测的条件概率。t_k是定义在边上的特征函数，称为 *转移特征* ，依赖于当前和前一个位置，s_l是定义在结点上的特征函数，称为 *状态特征* ，依赖于当前位置。t_k,s_l都依赖于位置，是局部特征函数。通常，特征函数的取值为1或0：当满足条件取1，否则为0。条件随机场完全由特征函数t_k,s_l及权重λ_k,μ_l决定。

条件随机场的简化形式

条件随机场还可以由简化形式表示。注意到上式中同一特征在各个位置都有定义，可以对同一特征在各个位置求和，将局部特征函数转化为一个全局特征函数，这样就可以将条件随机场写成权重向量和特征向量的内积形式。

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个转移特征，$K=K_1+K_2$，记

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

然后，对转移与状态特征在各个位置i求和，记作

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

用w_k表示特征$f_k(y,x)$的权值，即

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

于是，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表示权重向量，即$w=(w_1,w_2,\cdots,w_K)^T$，以$F(y,x)$表示全局特征向量，即$F(y,x)=(f_1,f_2,\cdots,f_K)^T$，则条件随机场可以写成向量的内积：

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

条件随机场的矩阵形式

条件随机场还可以由矩阵表示。假设$P_w(y|x)$是由上式给出的线性链条件随机场，引进特殊的起点的终点状态标记$y_0=start, y_{n+1}=stop$，这时$P_w(y|x)$可以通过矩阵形式表示。

对观测序列x的每一个位置$i=1,\cdots,n+1$，定义一个m阶矩阵(m是标记y_i取值的个数)：

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

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

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

注意，$y_0=start, y_{n+1}=stop$表示开始状态与终止状态，规范化因子$Z_w(x)$是以start为起点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。

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$P(Y_i=y_i|x), P(Y_{i-1}=y_{i-1}|x)$以及相应的数学期望的问题。

前向-后向算法

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

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

递推公式为

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

又可以表示为

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

$\alpha_i(y_1|x)$表示在位置i的标记是y_i且到位置i的前部分标记序列的非规范化概率，y_i可取的值有m个，所以$\alpha_i(x)$是m维向量。

同样，对每个指标$i=0,1,\cdots,n+1$，定义后向向量$\beta_i(x)$：

$$\beta_{n+1}(y_{n+1}|x)=\begin{cases}1, & y_{n+1}=stop\\ 0, & else\end{cases}$$
$$\beta_i(y_i|x)=[M_i(y_i,y_{i+1}|x)]\beta_{i+1}(y_{i+1}|x)$$
$$\beta_i(x)=M_{i+1}(x)\beta_{i+1}(x)$$

$\beta_i(y_i|x)$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。

于是可得：

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

概率计算

按照前后-后向向量的定义，很容易计算标记序列在位置i是标记y_i的条件概率和在位置$i-1$与i是标记y_{i-1}和y_i的条件概率：

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

期望值的计算

利用前向-后向向量，可以计算特征函数关于联合分布$P(X,Y)$和条件分布$P(Y|X)$的数学期望。

特征函数f_k关于条件分布$P(Y|X)$的数学期望是

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

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

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

对于转移特征，将上式的f_k换成t_k，对于状态特征，换成s_l。

现在对于给定的观测序列与标记序列，可以通过一次前向扫描计算$\alpha_i,Z(x)$，通过一次后向扫描计算β_i，从而计算所有的概率及期望。

条件随机场的学习算法

现在讨论给定训练数据集估计条件随机场模型参数的问题，即条件随机场的学习问题。条件随机场模型实际上是定义在时序数据上的对数线性模型，其学习方法包括极大似然估计和正则化的极大似然估计。具体的优化方法有改进的迭代尺度算法IIS、梯度下降法和拟牛顿法。

改进的迭代尺度算法

已知训练数据集，由此可知经验概率分布$\tilde{P}(X,Y)$，可以通过极大化训练数据的对数似然函数来求模型参数。

训练数据的对数似然为

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

IIS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$w=(w_1,\cdots,w_K)^T$，向量的增量为δ，更新参数向量为$w+\delta$。在每步迭代过程中，IIS通过依次求解下两式，得到δ。

关于转移特征t_k的更新方程为

$$
\begin{gathered}
E_{\tilde{P}}[t_k]=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n+1}t_k(y_{i-1},y_i,x,i) \\
=\sum_{x,y}\tilde{P}(x)P(y|x)\sum_{i=1}^{n+1}t_k(y_{i-1},y_i,x,i)\exp(\delta_k T(x,y)) \\
k=1,\cdots,K_1
\end{gathered}
$$

关于状态特征s_l的更新方程为

$$
\begin{gathered}
E_{\tilde{P}}[s_l]=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n+1}s_l(y_i,x,i) \\
=\sum_{x,y}\tilde{P}(x)P(y|x)\sum_{i=1}^{n}s_l(y_i,x,i)\exp(\delta_{K_1+l}T(x,y)) \\
l=1,\cdots,K_2
\end{gathered}
$$

这里，$T(x,y)$是在数据(x,y)中出现所有特征数的总和：

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

算法(条件随机场模型学习的IIS算法)：

1. 对所有的$k=1,2,\cdots,K$，取初值$w_k=0$
2. 对每一个k：
   1. 当$k=1,2,\cdots,K_1$，令δ_k是方程

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

      的解；

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

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

      的解。
   2. 更新$w_k=w_k+\delta_k$
3. 如果不是所有的w_k都收敛，重复2

$T(x,y)$表示数据(x,y)中的特征总数，对不同的数据(x,y)取值可能不同。为了处理这个问题，定义松弛特征

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

其中S是一个常数，选择足够大的常数能够使得对训练数据集的所有数据$(x,y),s(x,y)\geq 0$成立。这时特征总是可取S。

对于转移特征t_k，δ_k的更新方程是

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

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

其中，

$$E_p[t_k]=\sum_x\tilde{P}(x)\sum_{i=1}^{n+1}\sum_{y_{i-1},y_i}t_k(y_{i-1},y_i,x,i)\frac{\alpha_{i-1}^T(y_{i-1}|x)M_i(y_{i-1},y_i|x)\beta_i(y_i|x)}{Z(x)}$$

同样，对于状态特征s_l，δ_k的更新方程为

$$\sum_{x,y}\tilde{P}(x)P(y|x)\sum_{i=1}^{n+1}s_l(y_i,x,i)\exp(\delta_{K_1+l}S)=E_{\tilde{P}}[s_l]$$

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

其中，

$$E_p[s_l]=\sum_x\tilde{P}(x)\sum_{i=1}^{n}\sum_{y_i}s_l(y_i,x,i)\frac{\alpha_i^T(y_i|x)\beta_i(y_i|x)}{Z(x)}$$

以上算法称为算法S。需要使常数S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-后向算法，可以很容易计算$T(x)=t$。

这时，关于转移特征参数的更新方程可以写成：

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

这里，$a_{k,t}$是特征t_k的期望值，$\delta_k=\log\beta_k$。β_k是上述多项式方程的唯一实根，可以用牛顿法求得。

同样，对于状态特征的参数更新方程可以写成：

$$\begin{aligned}E_{\tilde{P}}[s_l]&=\sum_{x,y}\tilde{P}(x)P(y|x)\sum_{i=1}^{n}s_l(y_i,x,i)\exp(\delta_{K_1+l}T(x))\\&=\sum_x\tilde{P}(x)\sum_y P(y|x)\sum_{i=1}^{n}s_l(y_i,x,i)\exp(\delta_{K_1+l}T(x))\\&=\sum_x\tilde{P}(x)b_{l,t}\exp(\delta_k\cdot t)\\&=\sum_{t=0}^{T_{max}}b_{l,t}\gamma_l^t\end{aligned}$$

这里，$b_{l,t}$是特征s_l的期望值，$\delta_l=\log\gamma_l$，γ_l是上述多项式方程的唯一实根，也可以用牛顿法求得。

拟牛顿法

CRF模型学习还可以应用牛顿法或拟牛顿法。对于CRF模型

$$P_w(y|x)=\frac{\exp\left(\sum\limits_{i=1}^{n}w_i f_i(x,y)\right)}{\sum_y\exp\left(\sum\limits_{i=1}^{n}w_i f_i(x,y)\right)}$$

学习的优化目标函数是

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

其梯度函数是

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

算法(CRF模型学习的BFGS算法)：

1. 选定初始点w^o，取B_0为正定对称矩阵，令$k=0$
2. 计算$g_k=g(w^k)$。若$g_k=0$，则停止计算，否则到3
3. 由$B_k p_k=-g_k$求出p_k
4. 一维搜索：求λ_k使得

$$f(w^k+\lambda_k p_k)=\min_{\lambda\geq 0}f(w^k+\lambda p_k)$$

5. 令$w^{k+1}=w^k+\lambda_k p_k$
6. 计算$g_{k+1}=g(w^{k+1})$，若$g_{k+1}=0$，则停止计算；否则，按下式求出B_{k+1}：

$$B_{k+1}=B_k+\frac{y_k y_k^T}{y_k^T\delta_k}-\frac{B_k\delta_k\delta_k^T B_k}{\delta_k^T B_k\delta_k}$$

   其中，

$$y_k=g_{k+1}-g_k,\delta_k=w^{k+1}-w^k$$

7. 令$k=k+1$，转3

条件随机场的预测算法

CRF的预测问题是给定条件随机场$P(Y|X)$和输入序列，求概率最大的输出序列y^*。我们使用维特比算法。

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

于是，CRF的预测问题成为求非规范化概率最大的最优路径问题

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

这里，路径表示标记序列。其中

$$
\begin{aligned}
w &= (w_1, \cdots, w_K)^T \\
F(y, x) &= (f_1, \cdots, f_K)^T \\
f_k(y, x) &= \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i), k = 1, \cdots, K
\end{aligned}
$$

这时只需计算非规范化概率，而不必计算概率，可以大大提高效率。为此，将上式写成如下形式：

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

其中

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

是局部特征向量。

下面叙述维特比算法。首先求出位置1的各个标记$j = 1, \cdots, m$的非规范化概率

$$
\delta_1(j) = w \cdot F_1(y_0 = start, y_1 = j, x), j = 1, \cdots, m
$$

由递推公式，求出到位置i的各个标记$l = 1, \cdots, m$的非规范化概率的最大值，同时记录非规范化概率最大值的路径

$$
\begin{aligned}
\delta_i(l) &= \max_{1 \leq j \leq m} \{\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)\}, l = 1, \cdots, m \\
\Psi_i(l) &= \arg\max_{1 \leq j \leq m} \{\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)\}, l = 1, \cdots, m
\end{aligned}
$$

直到$i = n$终止。这时求得非规范化概率的最大值为

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

及最优路径的终点

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

由此最优路径终点返回，

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

求得最优路径$y^* = (y_1^*, \cdots, y_n^*)^T$。

算法(CRF预测的维特比算法)：

1. 初始化

$$
\delta_1(j) = w \cdot F_1(y_0 = start, y_1 = j, x), j = 1, \cdots, m
$$

2. 递推，对$i = 1, \cdots, n$

$$
\begin{aligned}
\delta_i(l) &= \max_{1 \leq j \leq m} \{\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)\}, l = 1, \cdots, m \\
\Psi_i(l) &= \arg\max_{1 \leq j \leq m} \{\delta_{i-1}(j) + w \cdot F_i(y_{i-1} = j, y_i = l, x)\}, l = 1, \cdots, m
\end{aligned}
$$

3. 终止

$$
\begin{aligned}
\max_y (w \cdot F(y, x)) &= \max_{1 \leq j \leq m} \delta_n(j) \\
y_n^* &= \arg\max_{1 \leq j \leq m} \delta_n(j)
\end{aligned}
$$

4. 返回路径

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

小结

1. 概率无向图模型是由无向图表示的联合概率分布。无向图上的结点之间的连接关系表示了联合分布的随机变量集合之间的条件独立性，即马尔科夫性。因此，概率无向图模型也称为马尔科夫随机场。

   概率无向图模型的联合概率分布可以分解为无向图最大团上的正值函数的形式。
2. 条件随机场是给定输入随机变量X条件下，给出随机变量Y的条件概率分布模型，其形式为参数化的对数线性模型。CRF的最大特点是假设输出变量之间的联合概率分布构成概率无向图模型。CRF是判别模型。
3. 线性链条件随机场是定义在观测序列与标记序列上的CRF。线性链CRF一般表示为给定观测序列条件下的标记序列的条件概率分布，由参数化的对数线性模型表示。
4. 线性链CRF概率计算通常利用前向-后向算法。
5. CRF的学习方法通常是极大似然估计方法，即在给定训练集下，通过极大化训练数据的对数似然估计参数。具体的方法是IIS，梯度下降，拟牛顿法。
6. 线性链CRF的一个重要应用是标注。维特比算法是给定观测序列求条件概率最大的标记序列的方法。