条件随机场

2021-09-26

1. 条件随机场概述
2. 概率无向图模型
   2. 1. 基本概念
   2. 2. 三种马尔可夫性
   2. 3. 概率无向图模型定义
   2. 4. 概率无向图模型的因子分解
3. 条件随机场表示
   3. 1. 条件随机场定义
   3. 2. 线性链条件随机场定义
   3. 3. 条件随机场的参数化形式
   3. 4. 条件随机场的简化形式
   3. 5. 条件随机场的矩阵形式
4. 条件随机场概率计算问题
   4. 1. 前向-后向算法
   4. 2. 概率计算
   4. 3. 期望值的计算
5. 条件随机场的学习算法
   5. 1. 改进的迭代尺度法
   5. 2. 拟牛顿法
6. 条件随机场的预测算法
7. 参考文档

1. 条件随机场概述

条件随机场（conditional random field, CRF）是一种**条件概率分布模型**，其特点是假设输出随机变量构成马尔可夫随机场。对于线性链条件随机场，其问题变成了由输入序列对输出序列预测的判别模型，形式为对数线性模型，学习方法通常是极大似然估计或正则化的极大似然估计。

2. 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。

补充：有向图则对应贝叶斯网络。

2.1. 基本概念

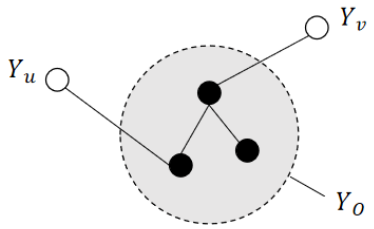
图是由结点及连接结点的边组成的集合。结点和边分别记作v和e，结点和边的集合分别记作V和E，图记作$G=(V,E)$。无向图是指边没有方向的图。

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。概率分布$P(Y)$由无向图$G=(V,E)$表示，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表示随机变量之间的概率依赖关系。

2.2. 三种马尔可夫性

给定一个联合概率分布$P(Y)$和表示该模型的无向图G。对无向图表示的随机变量之间存在的马尔可夫性进行定义，马尔可夫性包括成对马尔可夫性、局部马尔可夫性、全局马尔可夫性，并且，这三种独立性是等价的。

1. 成对马尔可夫性

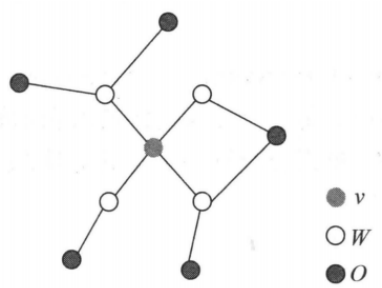


成对马尔可夫性

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。其它所有结点为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：

$$P(Y_u, Y_v \mid Y_O) = P(Y_u \mid Y_O) P(Y_v \mid Y_O) \tag{47}$$

2. 局部马尔可夫性



局部马尔可夫性

设$v \in V$是无向图G中任意一个结点，W是与v有边连接所有结点，O是v和W以外的其它所有结点。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。局部马尔可夫性是指给定随机变量组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：

$$\begin{aligned} P(Y_v, Y_O \mid Y_W) &= P(Y_v \mid Y_W) P(Y_O \mid Y_W) \\ &= P(Y_v \mid Y_O, Y_W) P(Y_O \mid Y_W) \end{aligned} \tag{48}$$

在$P(Y_O \mid Y_W) > 0$时，等价地：

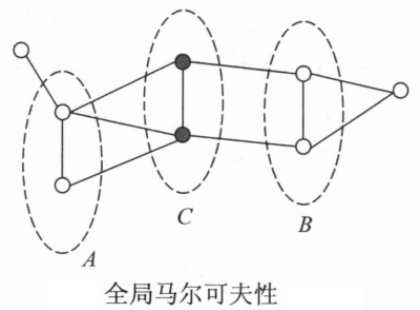
$$P(Y_v \mid Y_W) = P(Y_v \mid Y_O, Y_W) \tag{49}$$

该公式可以概括为：

$$P(Y_v \mid Y_Q, v \sim Q) = P(Y_v \mid Y_S, v \neq S) \tag{50}$$

其中，$v \sim Q$表示所有与v连接的点，$v \neq S$表示除了v以外的其它点

3. 全局马尔可夫性



全局马尔可夫性

设结点集合A，B是在无向图G中被结点集合C分开的任意结点集合。结点集合A，B和C所对应的随机变量组分别是Y_A，Y_B和Y_C。全局马尔可夫性是指给定随机变量组Y_C条件下随机变量组Y_A和Y_B是条件独立的，即

$$P(Y_A, Y_B \mid Y_C) = P(Y_A \mid Y_C) P(Y_B \mid Y_C) \tag{51}$$

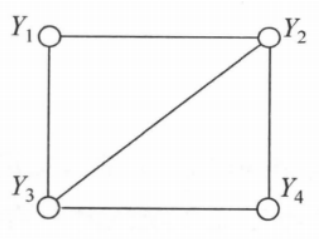
2.3. 概率无向图模型定义

定义：（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$P(Y)$为**概率无向图模型**，或**马尔可夫随机场**。

简记：概率无向图模型是$P(Y)$，$P(Y)$可以由无向图表示，且无向图结点间的概率满足马尔可夫性。

2.4. 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连接的结点子集称为团。若C是无向图G中的一个团，并且不能再加进任何一个G的结点使其称为一个更大的团，则称此C为最大团。



无向图的团和最大团

如上图所示的无向图，最大团为：$\{Y_1, Y_2, Y_3\}$和$\{Y_2, Y_3, Y_4\}$。

定理：$(Hammersley-Clifford$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：

$$P(Y)=\frac{1}{Z}\prod_{C}\Psi_C\left(Y_C\right)$$
$$Z=\sum_{Y}\prod_{C}\Psi_C\left(Y_C\right) \tag{52}$$

其中，C是无向图的最大团，Y_C是C中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。

上述定理表明，**概率无向图模型的联合概率分布**可以表示为其**最大团上的随机变量的函数的乘积形式。**

3. 条件随机场表示

3.1. 条件随机场定义

条件随机场（conditional random field)是给定随机变量X条件下，随机变量Y的马尔可夫随机场。

定义：(条件随机场) 设X和Y是随机变量，$P(Y \mid X)$是给定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$G=(V, E)$表示的马尔可夫随机场，即：

$$P\left(Y_v \mid X, Y_w, w \neq v\right)=P\left(Y_v \mid X, Y_w, w \sim v\right) \tag{53}$$

对任意结点v成立，则称条件概率分布$P(Y \mid X)$为条件随机场。式中$w \sim v$表示在图$G=(V, E)$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，Y_v、Y_u、Y_w为结点v、u、w对应的随机变量。

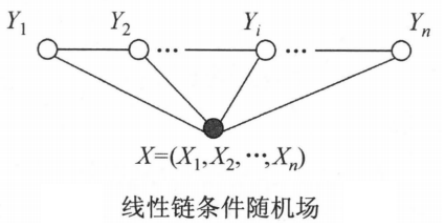
注意：在定义中并没有要求X和Y具有相同的结构。

3.2. 线性链条件随机场定义

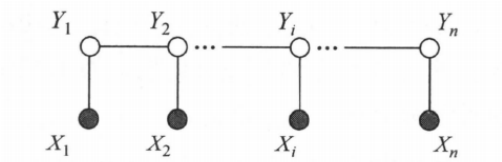
现实中，一般假设条件随机场中X和Y具有相同结构。现假设无向图为下图所示的线性链的情况，即图的定义为：

$$G=(V=\{1,2,\cdots,n\}, E=\{(i, i+1)\}), \quad i=1,2,\cdots,n-1 \tag{54}$$

在此情况下，$X=(X_1, X_2, \cdots, X_n), Y=(Y_1, Y_2, \cdots, Y_n)$，最大团是相邻两个结点的集合。



线性链条件随机场



X 和 Y 有相同的图结构的线性链条件随机场

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$P(Y \mid X)$中，Y是输出变量，表示**预测出的标记序列**，X是输入变量，表示**需要标注的观测序列。**例如，$X=\{$我爱中国$\}$，$Y=\{$名词，动词，名词$\}$。

定义：(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$P(Y\mid X)$构成条件随机场，即满足马尔可夫性：

$$\begin{gathered}P\left(Y_i\mid X,Y_1,\cdots,Y_{i-1},Y_{i+1},\cdots,Y_n\right)=P\left(Y_i\mid X,Y_{i-1},Y_{i+1}\right)\\i=1,2,\cdots,n(\text{ 在 }i=1\text{ 和 }n\text{ 时只考虑单边 })\end{gathered}\tag{55}$$

则称$P(Y\mid X)$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。

3.3. 条件随机场的参数化形式

根据$Hammersley-Clifford$定理，可以给出线性链条件随机场的因子分解式，各因子是定义在相邻两个结点（最大团）上的势函数。

定理：(线性链条件随机场的参数化形式) 设$P(Y\mid X)$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：

$$\begin{aligned}P(y\mid x)&=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}\mu_l s_l\left(y_i,x,i\right)\right)\\Z(x)&=\sum_y\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}\mu_l s_l\left(y_i,x,i\right)\right)\end{aligned}\tag{56}$$

其中，t_k是定义在边上的特征函数，称为转移特征；s_l是定义在结点上的特征函数，称为状态特征；t_k和s_l都依赖于位置，是局部特征函数，通常取值为0或1；λ_k和μ_l是对应的权值；$Z(x)$是规范化因子，求和是在所有可能的输出序列上进行的。

3.4. 条件随机场的简化形式

在条件随机场的参数化形式中，同一特征在各个位置都有定义，可以对同一个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值向量和特征向量的内积的形式，即条件随机场的简化形式。

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个状态特征，$K=K_1+K_2$，记：

$$f_k\left(y_{i-1},y_i,x,i\right)=\begin{cases}t_k\left(y_{i-1},y_i,x,i\right), & k=1,2,\cdots,K_1\\s_l\left(y_i,x,i\right), & k=K_1+l;l=1,2,\cdots,K_2\end{cases}\tag{57}$$

然后，对转移和状态特征在各个位置i求和，得到**全局特征函数**，记作：

$$f_k(y,x)=\sum_{i=1}^n f_k\left(y_{i-1},y_i,x,i\right),\quad k=1,2,\cdots,K\tag{58}$$

用w_k表示特征$f_k(y,x)$的权值，得到**全局特征权值**，即：

$$w_k=\begin{cases}\lambda_k, & k=1,2,\cdots,K_1\\\mu_l, & k=K_1+l;l=1,2,\cdots,K_2\end{cases}\tag{59}$$

于是，条件随机场(10)可表示为：

$$\begin{aligned}P(y\mid x)&=\frac{1}{Z(x)}\exp\sum_{k=1}^K w_k f_k(y,x)\\Z(x)&=\sum_y\exp\sum_{k=1}^K w_k f_k(y,x)\end{aligned}\tag{60}$$

若以$F(y,x)$表示全局特征向量，即：

$$F(y,x)=\left(f_1(y,x),f_2(y,x),\cdots,f_K(y,x)\right)^{\mathrm{T}}\tag{61}$$

以w表示权值向量，即：

$$w=\left(w_1,w_2,\cdots,w_K\right)^T\tag{62}$$

那么，条件随机场可以写成向量w与$F(y,x)$的内积的形式：

$$\begin{aligned}P_w(y\mid x)&=\frac{\exp\left(w\cdot F(y,x)\right)}{Z_w(x)}\\Z_w(x)&=\sum_y\exp\left(w\cdot F(y,x)\right)\end{aligned}\tag{63}$$

3.5. 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$P_w(y \mid x)$是由式(14)给出的线性链条件随机场，表示对给定观测序列x，相应的标记序列y的条件概率。

对每个标记序列引进特殊的起点和终点状态标记$y_0 = start$和$y_{n+1} = stop$，这时标注序列的概率$P_w(y \mid x)$可以通过矩阵形式表示并有效计算。

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**：

$$M_i(x) = \left[M_i\left(y_{i-1}, y_i \mid x\right)\right]_{m \times m} \tag{64}$$

矩阵随机变量的元素为：

$$\begin{aligned} M_i\left(y_{i-1}, y_i \mid x\right) &= \exp\left(W_i\left(y_{i-1}, y_i \mid x\right)\right) \\ W_i\left(y_{i-1}, y_i \mid x\right) &= \sum_{k=1}^{K} w_k f_k\left(y_{i-1}, y_i, x, i\right) \end{aligned} \tag{65}$$

其中，f_k和w_k分别由式(12)和式(13)给出，y_{i-1}和y_i是标记随机变量Y_{i-1}和Y_i的取值。

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该序列$n+1$个矩阵的**适当元素**的乘积$\prod_{i=1}^{n+1} M_i\left(y_{i-1}, y_i \mid x\right)$表示。因此，条件概率$P_w(y \mid x)$为：

$$\begin{aligned} P_w(y \mid x) &= \frac{1}{Z_w(x)} \prod_{i=1}^{n+1} M_i\left(y_{i-1}, y_i \mid x\right) \\ &= \frac{1}{Z_w(x)} \prod_{i=1}^{n+1} \exp\left(W_i\left(y_{i-1}, y_i \mid x\right)\right) \\ &= \frac{1}{Z_w(x)} \prod_{i=1}^{n+1} \exp\left(\sum_{k=1}^{K} w_k f_k\left(y_{i-1}, y_i, x, i\right)\right) \\ &= \frac{1}{Z_w(x)} \exp\left(\sum_{i=1}^{n+1} \sum_{k=1}^{K} w_k f_k\left(y_{i-1}, y_i, x, i\right)\right) \\ \text{其中，} Z_w(x) &= \left[M_1(x) M_2(x) \cdots M_{n+1}(x)\right]_{\text{start,stop}} \end{aligned} \tag{66}$$

可以看到，矩阵形式展开后与参数化形式或简化形式是一致的。

4. 条件随机场概率计算问题

4.1. 前向-后向算法

1. 前向向量

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

$$\alpha_0(y \mid x) = \begin{cases} 1, & y = \text{ start} \\ 0, & \text{否则} \end{cases} \tag{67}$$

递推公式为：

$$\begin{aligned} \alpha_i^{\mathrm{T}}\left(y_i \mid x\right) &= \alpha_{i-1}^{\mathrm{T}}\left(y_{i-1} \mid x\right)\left[M_i\left(y_{i-1}, y_i \mid x\right)\right] \\ \text{其中，} i &= 1, 2, \cdots, n+1 \end{aligned} \tag{68}$$

又可表示为：

$$\alpha_i^{\mathrm{T}}(x) = \alpha_{i-1}^{\mathrm{T}}(x) M_i(x) \tag{69}$$

$\alpha_i(y_i \mid x)$表示在位置i的标记是y_i并且从1到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维列向量。

2. 后向向量

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

$$\beta_{n+1}\left(y_{n+1} \mid x\right) = \begin{cases} 1, & y_{n+1} = \text{ stop} \\ 0, & \text{否则} \end{cases} \tag{70}$$

递推公式为：

$$\beta_i\left(y_i \mid x\right) = \left[M_{i+1}\left(y_i, y_{i+1} \mid x\right)\right] \beta_{i+1}\left(y_{i+1} \mid x\right) \tag{71}$$

又可表示为：

$$\beta_i(x) = M_{i+1}(x) \beta_{i+1}(x) \tag{72}$$

$\beta_i(y_i \mid x)$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。

4.2. 概率计算

根据前向-后向向量的定义，可以得出如下概率：

$$
\begin{aligned}
P\left(Y_i=y_i \mid x\right) &= \frac{\alpha_i^{\mathrm{T}}\left(y_i \mid x\right)\beta_i\left(y_i \mid x\right)}{Z(x)} \\
P\left(Y_{i-1}=y_{i-1}, Y_i=y_i \mid x\right) &= \frac{\alpha_{i-1}^{\mathrm{T}}\left(y_{i-1} \mid x\right)M_i\left(y_{i-1}, y_i \mid x\right)\beta_i\left(y_i \mid x\right)}{Z(x)} \\
&\text{其中，} Z(x)=\alpha_n^{\mathrm{T}}(x)\mathbf{1}=\mathbf{1}\beta_1(x) \\
&\mathbf{1}\text{是元素均为}1\text{的}m\text{维列向量}
\end{aligned}
\tag{73}
$$

4.3. 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$P(Y \mid X)$的数学期望。

1. 特征函数f_k关于条件分布$P(Y \mid X)$的数学期望

$$
\begin{aligned}
E_{P(Y|X)}\left[f_k\right] &= \sum_y P(y \mid x) f_k(y,x) \\
&= \sum_{i=1}^{n+1}\sum_{y_{i-1}y_i} f_k\left(y_{i-1}, y_i, x, i\right)\frac{\alpha_{i-1}^{\mathrm{T}}\left(y_{i-1} \mid x\right)M_i\left(y_{i-1}, y_i \mid x\right)\beta_i\left(y_i \mid x\right)}{Z(x)} \\
Z(x) &= \alpha_n^{\mathrm{T}}(x)\mathbf{1} \\
&\text{其中，} k=1,2,\cdots,K
\end{aligned}
\tag{74}
$$

2. 特征函数f_k关于联合分布$P(X,Y)$的数学期望

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

$$
\begin{aligned}
E_{P(X,Y)}\left[f_k\right] &= \sum_{x,y} P(x,y)\sum_{i=1}^{n+1} f_k\left(y_{i-1}, y_i, x, i\right) \\
&= \sum_x \tilde{P}(x)\sum_y P(y \mid x)\sum_{i=1}^{n+1} f_k\left(y_{i-1}, y_i, x, i\right) \\
&= \sum_x \tilde{P}(x)\sum_{i=1}^{n+1}\sum_{y_{i-1}y_i} f_k\left(y_{i-1}, y_i, x, i\right)\frac{\alpha_{i-1}^{\mathrm{T}}\left(y_{i-1} \mid x\right)M_i\left(y_{i-1}, y_i \mid x\right)\beta_i\left(y_i \mid x\right)}{Z(x)} \\
Z(x) &= \alpha_n^{\mathrm{T}}(x)\mathbf{1} \\
&\text{其中，} k=1,2,\cdots,K
\end{aligned}
\tag{75}
$$

5. 条件随机场的学习算法

5.1. 改进的迭代尺度法

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函数为：

$$
L(w)=L_{\tilde{P}}\left(P_w\right)=\log\prod_{x,y}P_w(y \mid x)^{\tilde{P}(x,y)}=\sum_{x,y}\tilde{P}(x,y)\log P_w(y \mid x)
\tag{76}
$$

当P_w是一个由式(14)给出的条件随机场模型时，对数似然函数为：

$$
\begin{aligned}
L(w) &= \sum_{x,y}\tilde{P}(x,y)\log P_w(y \mid x) \\
&= \sum_{x,y}\left[\tilde{P}(x,y)\sum_{k=1}^{K} w_k f_k(y,x)-\tilde{P}(x,y)\log Z_w(x)\right] \\
&= \sum_{j=1}^{N}\sum_{k=1}^{K} w_k f_k\left(y_j, x_j\right)-\sum_{j=1}^{N}\log Z_w\left(x_j\right)
\end{aligned}
\tag{77}
$$

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$\delta=(\delta_1,\delta_2,\cdots,\delta_K)$，更新参数向量为$w+\delta=(w_1+\delta_1,w+\delta_2,\cdots,w_K+\delta_K)^T$。在每步迭代过程中，改进的迭代尺度法通过依次求解式(32)和式(33)，得到$\delta=(\delta_1,\delta_2,\cdots,\delta_K)$。

1. 关于转移特征t_k的更新方程

$$
\begin{aligned}
E_{\tilde{P}}\left[t_k\right] &= \sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n+1} t_k\left(y_{i-1}, y_i, x, i\right) \\
&= \sum_{x,y}\tilde{P}(x)P(y \mid x)\sum_{i=1}^{n+1} t_k\left(y_{i-1}, y_i, x, i\right)\exp\left(\delta_k T(x,y)\right) \\
&\text{其中，} k=1.2.\cdots.K_1
\end{aligned}
\tag{78}
$$

2. 关于状态特征s_l的更新方程

$$\begin{aligned} E_{\tilde{P}}\left[s_l\right] &= \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{n+1} s_l\left(y_i, x, i\right) \\ &= \sum_{x,y} \tilde{P}(x) P(y \mid x) \sum_{i=1}^{n} s_l\left(y_i, x, i\right) \exp\left(\delta_{K_1+l} T(x,y)\right) \\ &\text{其中，} l = 1, 2, \cdots, K_2 \end{aligned} \tag{79}$$

这里，$T(x,y)$是在数据(x,y)中出现所有特征数的总和：

$$T(x,y) = \sum_k f_k(y,x) = \sum_{k=1}^{K} \sum_{i=1}^{n+1} f_k\left(y_{i-1}, y_i, x, i\right) \tag{80}$$

算法(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$\tilde{P}(x,y)$

输出：参数估计值$\hat{w}$；模型$P_{\hat{w}}$

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

(2)对每一 $k \in \{1, 2, \cdots, K\}$：

(a)当$K = 1, 2, \cdots, K_1$时，令δ_k是如下方程的解：

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

$$\text{解得：} \delta_k = \frac{1}{S} \log \frac{E_{\tilde{P}}\left[t_k\right]}{E_P\left[t_k\right]}$$

$$\text{其中，} E_P\left(t_k\right) = \sum_x \tilde{P}(x) \sum_{i=1}^{n+1} \sum_{y_{i-1}, y_i} t_k\left(y_{i-1}, y_i, x, i\right) \frac{\alpha_{i-1}^{\mathrm{T}}\left(y_{i-1} \mid x\right) M_i\left(y_{i-1}, y_i \mid x\right) \beta_i\left(y_i \mid x\right)}{Z(x)} \tag{81}$$

当$K = K_1 + l,\ l = 1, 2, \cdots, K_2$时，令$\delta_{K_1+l}$是如下方程的解：

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

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

$$\text{其中，} E_P\left(s_l\right) = \sum_x \tilde{P}(x) \sum_{i=1}^{n} \sum_{y_i} s_l\left(y_i, x, i\right) \frac{\alpha_i^{\mathrm{T}}\left(y_i \mid x\right) \beta_i\left(y_i \mid x\right)}{Z(x)} \tag{82}$$

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

(3)如果不是所有w_k都收敛，重复步骤(2)

在式(34)中，$T(x,y)$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。为了处理该问题，定义松弛特征：

$$s(x,y) = S - \sum_{i=1}^{n+1} \sum_{k=1}^{K} f_k\left(y_{i-1}, y_i, x, i\right) \tag{83}$$

式中S是一个常数。选择足够大的常数S使得对训练数据集的所有数据(x,y)，$s(x,y) \geqslant 0$成立。此时特征总数可取S。

5.2. 拟牛顿法

条件随机场的对数似然函数为：

$$L(w) = \sum_{x,y} \tilde{P}(x,y) \log P_w(y \mid x) \tag{84}$$

将条件随机场的简化形式带入得：

$$\begin{aligned} L(w) &= \sum_{x,y} \tilde{P}(x,y) \log \left[\frac{\exp \sum_{k=1}^{K} w_k f_k(y,x)}{Z_w(x)} \right] \\ &= \sum_{x,y} \tilde{P}(x,y) \left[\log \left(\exp \sum_{k=1}^{K} w_k f_k(y,x) \right) - \log Z_w(x) \right] \\ &= \sum_{x,y} \tilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \sum_{x,y} \tilde{P}(x,y) \log Z_w(x) \\ &= \sum \tilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \sum \sum \tilde{P}(x,y) \log Z_w(x) \end{aligned} \tag{85}$$

$$
\begin{aligned}
&\underset{x,y}{} \underset{k=1}{} \underset{x}{} \underset{y}{} \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \sum_{x} \log Z_w(x) \left(\sum_{y} \tilde{P}(x,y) \right) \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \sum_{x} \log Z_w(x) \tilde{P}(x) \\
&= \sum_{x,y} \tilde{P}(x,y) \sum_{k=1}^{K} w_k f_k(y,x) - \sum_{x} \tilde{P}(x) \log \sum_{y} \exp \sum_{k=1}^{K} w_k f_k(y,x)
\end{aligned}
$$

令$f(w)=-L(w)$，则最大化$L(w)$等价于最小化$f(w)$，所以，优化目标函数为：

$$
\min_{w \in \mathbf{R}^n} f(w) = \sum_{x} \tilde{P}(x) \log \sum_{y} \exp \left(\sum_{i=1}^{K} w_i f_i(x,y) \right) - \sum_{x,y} \tilde{P}(x,y) \sum_{i=1}^{K} w_i f_i(x,y) \tag{86}
$$

$f(w)$的梯度函数是：

$$
g(w) = \sum_{x,y} \tilde{P}(x) P_w(y \mid x) f(x,y) - E_{\tilde{P}}(f) \tag{87}
$$

算法 (条件随机场模型学习的**BFGS**算法)

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

输出：最优参数值$\hat{w}$；最优模型$P_{\hat{w}}(y \mid x)$

(1)选定初始点$w^{(0)}$，取$\mathbf{B}_0$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$\mathbf{B}_k p_k = -g_k$求出p_k

(4)一维搜索：求λ_k使得：

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

(5)置$w^{(k+1)} = w^{(k)} + \lambda_k p_k$

(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$\mathbf{B}_{k+1}$

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

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

(7)置$k=k+1$，转(3)

6. 条件随机场的预测算法

已知：

$$
\begin{aligned}
w &= (w_1, w_2, \cdots, w_K)^{\mathrm{T}} \\
F(y,x) &= (f_1(y,x), f_2(y,x), \cdots, f_K(y,x))^{\mathrm{T}} \\
f_k(y,x) &= \sum_{i=1}^{n} f_k\left(y_{i-1}, y_i, x, i\right), \quad k=1,2,\cdots,K
\end{aligned} \tag{88}
$$

根据条件随机场的矩阵形式，可得：

$$
\begin{aligned}
y^* &= \arg\max_{y} P_w(y \mid x) \\
&= \arg\max_{y} \frac{\exp\left(w \cdot F(y,x)\right)}{Z_w(x)} \\
&= \arg\max_{y} \exp\left(w \cdot F(y,x)\right) \\
&= \arg\max_{y} (w \cdot F(y,x))
\end{aligned} \tag{89}
$$

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：

$$
\max_{y} (w \cdot F(y,x)) \tag{90}
$$

根据定义可以推导：

$$
\begin{aligned}
w \cdot F(y,x) &= (w_1, w_2, \cdots, w_K)^{\mathrm{T}} \cdot (f_1(y,x), f_2(y,x), \cdots, f_K(y,x))^{\mathrm{T}} \\
&= (w_1, w_2, \cdots, w_K)^{\mathrm{T}} \cdot \left(\sum_{i=1}^{n} f_1\left(y_{i-1}, y_i, x, i\right), \sum_{i=1}^{n} f_2\left(y_{i-1}, y_i, x, i\right), \cdots, \sum_{i=1}^{n} f_K\left(y_{i-1}, y_i, x, i\right) \right)^{\mathrm{T}}
\end{aligned}
$$

$$
\begin{aligned}
&=\sum_{i=1}^{n}\left(w_1, w_2, \cdots, w_K\right)^{\mathrm{T}} \cdot\left(f_1\left(y_{i-1}, y_i, x, i\right), f_2\left(y_{i-1}, y_i, x, i\right), \cdots, f_K\left(y_{i-1}, y_i, x, i\right)\right)^{\mathrm{T}} \\
&=\sum_{i=1}^{n} w \cdot F_i\left(y_{i-1}, y_i, x, i\right)
\end{aligned} \tag{91}
$$

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

因此，优化目标可以写成如下形式：

$$
\max_{y} \sum_{i=1}^{n} w \cdot F_i\left(y_{i-1}, y_i, x\right)\text{其中，}F_i\left(y_{i-1}, y_i, x, i\right)=\left(f_1\left(y_{i-1}, y_i, x, i\right), f_2\left(y_{i-1}, y_i, x, i\right), \cdots, f_K\left(y_{i-1}, y_i, x, i\right)\right)^{\mathrm{T}} \tag{92}
$$

算法 (条件随机场的维特比算法)

输入：模型特征向量$F_i(y_{i-1}, y_i, x, i), i=1,2, \cdots, n$；权值向量$w$；观测序列$x=\left(x_1, x_2, \cdots, x_n\right)$

输出：最优路径$y^*=\left(y_1^*, y_2^*, \cdots, y_n^*\right)$

(1)初始化

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

(2)递推：对$i=2,3, \cdots, n$

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

$$
\Psi_i(l)=\arg \max_{1 \leqslant j \leqslant m}\left\{\delta_{i-1}(j)+w \cdot F_i\left(y_{i-1}=j, y_i=l, x\right)\right\}, \quad l=1,2, \cdots, m
$$

(3)终止

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

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

(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：$y^*=\left(y_1^*, y_2^*, \cdots, y_n^*\right)$

7. 参考文档

1. 《统计学习方法》啃书手册 | 第11章 条件随机场 - 知乎 (zhihu.com)
2. 一文读懂机器学习概率图模型（附示例和学习资源） - 云+社区 - 腾讯云 (tencent.com)