**8.条件随机场**

**8.1概率无向图模型**

**8.1.1模型的定义**

**图（graph）的定义**：由结点（node，记作v）及连接结点的边（edge，记作e）组成的集合。结点和边的集合分别记作V和E，图记作G = (V，E)。图分为无向图和有向图，无向图是边没有方向的图，它们将分解表示成一组函数；有向图是使用有向边的图，它们用条件概率分布表示分解。

**概率图模型（probabilistic graphical model）的定义**：由图表示的概率分布。设有联合概率分布P(Y)（Y∈У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∈V表示一个随机变量Yv，Y = (Yv)v∈V）；边e∈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。无向图表示的随机变量之间存在三个等价的定义：成对马尔可夫性（pairwise Markov property）、局部马尔可夫性（local Markov property）和全局马尔可夫性（global Markov property）。

**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点，结点u和v分别对应随机变量Yu和Yv，其他所有结点为O，对应的随机变量组是YO。成对马尔可夫性指给定随机变量组YO的条件下随机变量Yu和Yv是独立的，即

P(Yu，Yv | YO) = P(Yu | YO) P(Yv | YO)

**局部马尔可夫性**：设v∈V是无向图G中任意一个结点，W是与v有边连接的所有结点，O是v，W以外的其他所有结点。v表示随机变量Yv，W表示随机变量YW，O表示随机变量YO。局部马尔可夫性是指给定随机变量组YW的条件下随机变量Yv和随机变量组YO是独立的，即

P(Yv，YO | YW) = P(Yv | YW) P(YO | YW)

**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、C对应的随机变量组分别是YA，YB，YC。全局马尔可夫性指给定随机变量组YC条件下随机变量组YA和YB是条件独立的，即

P(YA，YB | YC) = P(YA | YC) P(YB | YC)

**概率无向图模型定义**：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（probability undirected graphical model）或马尔可夫随机场（Markov random field）。

概率无向图模型的最大特点就是**易于因子分解**，将联合概率分布进行因子分解即将整体的联合概率写成若干联合概率乘积的形式。

**8.1.2概率无向图模型的因子分解**

**团的定义**：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）。

**最大团的定义**：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）。

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**（factorization）：

给定概率无向图模型，设其无向图为G，C为G上的最大团，YC表示C对应的随机变量。那么概率无向图模型的联合概率分布P(Y)可写作图中所有最大团C上的函数ψC(YC)的乘积形式，即

其中Z是规范化因子（normalization factor）

规范化因子保证P(Y)构成一个概率分布。函数ψC(YC)称为势函数（potential function），这里要求势函数ψC(YC)是严格正的，通常定义为指数函数

ψC(YC) = exp{-E(YC)}

概率无向图模型的因子分解由Hammersley-Clifford定理来保证。

**8.2条件随机场的定义与形式**

**8.2.1条件随机场的定义**

条件随机场是给定输入随机变量X的条件下，输出随机变量Y的条件概率分布模型，其形式为参数化的对数线性模型。条件随机场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，即马尔可夫随机场。条件随机场是判别模型。

（1）**一般的条件随机场的定义**：

设X与Y是随机变量，P(Y | 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(Yv | X，Yw，w≠v) = P(Yv | X，Yw，w~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。式中w~v表示在图G = (V，E)中与结点v有边连接的所有结点w，w≠v表示结点v以外的所有结点，Yv与Yw为结点v，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，但在现实中一般假设X和Y具有相同的图结构。

（2）**线性链条条件随机场的定义**：

设X = (X1，X2，…，Xn)，Y = (Y1，Y2，…，Yn)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(Yi | X，Y1，…，Yi-1，Yi+1，…，Yn) = P(Yi | X，Yi-1，Yi+1)

i = 1，2，…，n（在i = 1和n时只考虑单边）

线性链条件随机场可用于标注等问题。这时在条件概率模型P(Y | X)中，X是输入变量，表示需要标注的观测序列，Y是输出变量，表示标记序列（也称状态序列）。

**8.2.2（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

其中Z(x)是规范化因子，是在所有可能的输出序列上求和

tk称为转移特征，是定义在边上的特征函数，依赖于当前和前一个位置；sl称为状态特征，是定义在结点上的特征函数，依赖于当前位置。tk和sl都依赖于位置，是局部特征函数，通常取值为1或0，当满足特征条件时取值为1，否则为0。λk和μl是tk和sl对应的权值。条件随机场完全由特征函数tk、sl和对应的权值λk、μl确定。

（线性链）条件随机场还可以由简化形式和矩阵形式表示：

**（1）简化形式**

条件随机场的形式中同一个特征各个位置都有定义，所以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，即将条件随机场写成权值向量和特征向量的内积形式：

设有K1个转移特征，K2个状态特征，K = K1 + K2，记

然后对转移和状态特征在各个位置i求和，记作

再用wk表示特征fk(y，x)的权值，即

于是条件随机场的形式可以表示为

若以w表示权值向量

w = (w1，w2，…，wK)T

以F(y，x)表示全局特征向量

F(y，x) = (f1(y，x)，f2(y，x)，…，fK(y，x))T

则条件随机场可以写成向量w和F(y，x)的内积形式

其中

**（2）矩阵形式**

通过引进特殊的起点和终点状态标记y0 = start，yn+1 = stop，Pw(y | x)可以用矩阵形式表示。对观测序列x的每一个位置i = 1，2，…，n+1定义一个m阶矩阵（m是标记yi的取值个数）：

Mi(x) = [Mi (yi-1，yi | x)]

Mi (yi-1，yi | x) = exp(Wi (yi-1，yi | x))

于是，给定观测序列x，标记序列y的非规范化概率可以通过n+1个矩阵的乘积表示，因此条件概率Pw(y | x)可以表示为

其中规范化因子Zw(x)是n+1个矩阵的乘积的（start，stop）元素，是以y0=start为起点yn+1=stop为终点通过状态的所有路径y1 y2…yn的非规范化概率之和

Zw(x) = (M1(x) M2(x)…Mn+1(x))start,stop

**8.3三个基本问题**

和隐马尔科夫模型一样，条件随机场也有三个基本问题：概率计算问题（评估问题）、学习问题以及预测问题（解码问题）。

**8.3.1概率计算问题**

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(Yi = yi | x)，P(Yi-1 = yi-1，Yi = yi | x)以及相应的数学期望。

首先，和隐马尔科夫模型一样引入前向变量和后向变量。

a）**前向向量**：αi(yi | x)表示在位置i的标记是yi且到位置i前的部分的标记序列的非规范化概率，yi可取的值有m个，由这m个αi(yi | x)组成的m维列向量就是前向向量αi(x)。

当i = 0时

递推公式为

所以前向向量可表示为

b）**后向向量**：βi(yi | x)表示在位置i的标记为yi且从i+1到n的部分的标记序列的非规范化概率，由m个βi(yi | x)组成的m维列向量为后向向量βi(x)。

当i = n+1时

递推公式为

所以后向向量可表示为

由前向-后向向量的定义可得

其中**1**是元素均为1的m维列向量。

（1）概率计算

按照前向-后向向量的定义，计算标记序列在位置i的标记是yi的条件概率

在位置i-1与i是标记yi-1和yi的条件概率

其中

（2）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概率分布P(X，Y)和条件概率分布P(Y | X)的数学期望。

特征函数fk关于条件分布P(Y | X)的数学期望是

其中

假设经验分布为(X)，特征函数fk关于联合分布P(X，Y)的数学期望是

其中

fk可以是转移特征函数tk，也可以是状态特征函数sl。有了概率计算和特征期望计算的公式，对给定的观测序列x和标记序列y，就可以通过一次前向扫描计算αi及Z(x)，通过一次后向扫描计算βi，计算所有的概率和特征的期望。

**8.3.2学习问题**

（线性链）条件随机场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，具体的实现算法有：改进的迭代尺度法（IIS）、梯度下降法和拟牛顿法等。

**（1）改进的迭代尺度法**

给定训练数据集，条件随机场的目标函数对数似然函数为

因为条件随机场

代入对数似然函数可得

从最大熵模型的推导可知，改进的迭代尺度法是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。令在数据(x，y)中出现的所有特征数的总和为

条件随机场的特征函数fk(y，x)有转移和状态两个方程，更新方程分别为：

a）转移特征tk的更新方程为

b）状态特征sl的更新方程为

**条件随机场模型学习的改进的迭代尺度法**

输入：特征函数t1，t2，…，tK1，s1，s2，…，sK2；经验分布(x，y)；

输出：参数估计值；模型。

1）对所有k∈{1，2，…，K}，取初值wk = 0；

2）对每一k∈{1，2，…，K}：

（a）当k = 1，2，…，K1时，令δk是方程

的解；当k = K1 + l，l = 1，2，…，K2时，令δk是方程

的解。

（b）更新wk值：wk ← wk + δk；

3）如果不是所有的wk都收敛，重复步骤2）。

上述算法有一个问题，就是表示数据(x，y)的特征总数的T(x，y)对不同的数据(x，y)取值可能不同，为了解决这个问题，可以使用算法S和算法T。

算法S定义松弛特征

式中S是一个足够大的常数，使对训练数据集中的所有数据(x，y)都有s(x，y) ≥ 0，这时特征总数就可取S。于是根据转移特征tk，δk的更新方程

可推出

其中

同样根据状态特征sl，δk的更新方程

可退出

其中

算法S有一个比较大的缺陷，因为一般需要常数S足够大，导致每步迭代的增量向量会变大，收敛会变慢。算法T就尝试解决这个问题，对每个观测序列x计算其特征总数最大值T(x)：

计算T(x) = t，此时转移特征参数的更新方程为

其中是特征tk的期望值，。βk是上面的方程唯一的实根，可用牛顿法求解，从而求得δk。

状态函数参数的更新方程为

跟前面一样，是特征sl的期望值，。γl是上面的方程唯一的实根，也可用牛顿法求解，从而求得δl。

**（2）拟牛顿法**

对于条件随机场模型

学习的优化目标函数是

则目标函数的梯度函数是

**条件随机场学习的BFGS算法：**

输入：特征函数f1，f2，…，fk；经验分布(X，Y)；

输出：最优参数估计值；最优模型。

1）选定初始点w(0)，取**B**0为正定对称矩阵，置k = 0；

2）计算gk = g(w(k))。若gk = 0，则停止计算；否则转3）；

3）由Bkpk = -gk求出pk；

4）一维搜索：求λk使得

5）置w(k+1) = w(k) + λkpk；

6）计算gk+1 = g(w(k+1))，若gk = 0，则停止计算；否则按下式求出Bk+1：

其中yk = gk+1 - gk，δk = w(k+1) – w(k)。

7）置k = k + 1，转3）。

**8.3.3预测问题**

条件随机场的预测问题是给定条件随机场模型P(Y | X)和输入序列（观测序列）x，求解条件概率最大的输出序列（标记序列即状态序列）y\*，标记序列是对观测序列的标注。

由条件随机场的简化形式可得

于是条件随机场的预测问题就变成求非规范化概率最大的最优路径问题

路径表示标记序列（状态序列），其中

w = (w1，w2，…，wK)T

F(y，x) = (f1(y，x)，f2(y，x)，…，fK(y，x))T

所以条件随机场的预测算法也可以用维特比算法。

将上面的问题写成如下形式

其中Fi(yi-1，yi，x) = (f1(yi-1，yi，x，i)，f2(yi-1，yi，x，i)，…，fK(yi-1，yi，x，i))T是局部特征向量。

跟隐马尔科夫模型一样，首先定义维特比算法相关向量：a）到位置i的各个标记l = 1，2，…，m的非规范化概率的最大值δi(l)；b）到位置i的非规范化概率值最大的路径上位置i-1的标记ψi(l)。

位置为1时各个标记j = 1，2，…，m的非规范化概率：

于是δi(l)的递推公式：

同时ψi(l)的递推公式：

当i = n时递推终止，求得非规范化概率最大值为

此时最优路径的终点

然后由此最优路径终点返回

求得最优路径。

**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=(x1，x2，…，xn)；

输出：最优路径。

（1）初始化

（2）递推。对i = 2，3，…，n

（3）终止。

（4）回溯路径。

求得最优路径。