从逻辑回归到条件随机场模型

预备知识

1. **概率无向图模型（又称马尔可夫随机场）的联合概率公式**

   在马尔可夫随机场中，给定两个变量子集的分离集，则这两个变量子集条件独立，所以联合概率如下式所示：

$$P(y_1, ..., y_n) = \frac{1}{Z}\prod_C \psi_C(Yc) \tag{1.1}$$

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

　　其中：c为最大团，Y_C是C中的结点对应的随机变量，$\psi_C(Y_C)$是在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。

由此可见，马尔可夫随机场是生成式模型

2. **生成式模型与判别式模型**

生成式模型与判别式模型的本质区别在于模型中观测序列x和状态序列y之间的决定关系，前者假设y决定x，后者假设x决定y。所以，生成式模型对$p(x,y)$建模，判别式模型对$p(y|x)$建模。

3. **应用示例**

为了便于后面模型的应用，先假设有一批词性标注的语料。即：

- 若干条已分好词的句子
- 每个句子中词的词性（注意：同一个词可以有不同的词性，例如一打汽水，打是量词，而打灯笼，打是动词）。

可以看出，通常从句子中提取特征feature，句子中词的词性是待预测的label。

正式开始

逻辑回归LR

先说最大熵模型MaxEnt，最大熵模型是在满足约束条件的模型集合中选取熵最大的模型。

定义在条件概率$P(Y|X)$上的条件熵为：

$$H(P) = -\sum_x P(x) \sum_y P(y|x) logP(y|x) = -\sum_{x,y} \widetilde{P}(x) P(y|x) logP(y|x)$$

最大熵模型的学习等价于约束最优化问题：

$$\min_{p \in C} -H(P) = \sum_{x,y} \widetilde{P}(x) P(y|x) logP(y|x)$$

$$\text{s.t. } E_P(f_i) - E_{\widetilde{P}}(f_i) = 0, i = 1, 2, ..., n$$
$$\sum_y P(y|x) = 1$$

经过拉格朗日函数及对偶问题，最大熵模型的一般形式为：

$$P_w(y|x) = \frac{1}{Z_w(x)} exp(\sum_{i=1}^n w_i f_i(x, y)) \tag{2.1}$$

$$Z_w(x) = \sum_y exp(\sum_{i=1}^n w_i f_i(x, y))$$

LR的形式为：

$$P(Y = 1|x) = \frac{1}{Z} exp(w_k \cdot x)$$
$$P(Y = 0|x) = \frac{1}{Z} \tag{2.2}$$

其中，

$$Z = 1 + \sum_{k=1}^1 exp(w_k \cdot x)$$

对于给定数据集，$x = (x^{(1)}, .., x^{(n)})^T$定义有n个约束，有如下特征函数，对比式2.1、2.2，可以看出，最大熵模型就变成了LR

$$f_i(x, y) = \begin{cases} x^{(i)}, y = 1 \\ 0, y = 0 \end{cases}$$

使用MaxEnt对3.应用示例进行建模，第i个特征函数示例

$$f_i(x, y) = \begin{cases} 1, ify = [V] \ and \ x = \text{"喝"} \\ 0, \ otherwise \end{cases}$$

条件随机场CRF

HMM认为，在任意t时刻的状态只依赖于前一个时刻的状态，与其他时刻的状态和观测无关。这会有什么问题呢？以应用示例为例，当前词性只依赖于上一个词性吗？按照我们的经

验，会认为，当前词性不仅依赖于上一个词性，还依赖于当前词，甚至还依赖于句子的末尾符号（例如，英文句子末尾为问号，会影响首字符的词性），这时CRF就出场了。

CRF是给定随机变量X条件下，随机变量Y的马尔可夫随机场。例如，给定分好词的句子，输出相应的词性，这就是条件随机场。由于输出结果是线性的，所以称为线性链条件随机场，下面给出线性链条件随机场的定义：

设$X=(X_1,X_2,...,X_n),Y=(Y_1,Y_2,...,Y_n)$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：

$$P(Y_i|X,Y_1,..Y_{i-1}Y_{i+1},...,Y_n)=P(Y_i|X,Y_{i-1},Y_{i+1}) \tag{3.1}$$

根据式1.1和3.1，可得：

$$P(y_1,y_2,...,y_n|x)=\frac{1}{Z}\prod_i\psi(y_i,y_i-1,x) \tag{3.2}$$

在线性链条件随机场有两种特征函数：

- 转移特征函数t_k，定义在边上
- 状态特征函数s_l，定义在节点上

势函数通常定义为指数函数，将两种特征函数代入式3.2得：

$$P(y|x)=\frac{1}{Z(x)}exp(\sum_{i,k}\lambda_k t_k(y_{i-1},y_i,x,i)+\sum_{i,l}\mu_l s_l(y_i,x,i)) \tag{3.3}$$

观察式3.3发现，同一个特征函数在各个位置都有定义。使用$f_m(y_{i-1},y_i,x,i)$统一表示$t_k(y_i-1,y_i,x,i),s_l(y_i,x,i)$，对特征函数在各个位置$i$求和，记作

$$f_m(y,x)=\sum_{i=1}^{n}f(y_{i-1},y_i,x,i),m=1,2,...,k+l \tag{3.4}$$

使用w_m统一表示λ_k,μ_l，将式3.4代入3.3得

$$\begin{gathered}P(y|x)=\frac{1}{Z(x)}exp\sum_{m=1}^{k+l}w_m f_m(y,x)\\ Z(x)=\sum_y exp\sum_{m=1}^{k+l}w_m f_m(y,x)\end{gathered} \tag{3.5}$$

对比式3.5、式2.1发现是一样的，Linear-Chain CRF的f_m是对特征函数在各个位置i的求和，所以CRF是最大熵模型的序列化，而LR是最大熵模型的特殊形式，所以在《统计自然语言处理(第2版)》里的图6-2显示，LR+sequence->Linear-Chain CRFs。

使用Linear-Chain CRFs对3.应用示例进行建模，第k个和第l个特征函数具体例子：

$$t_k(y_i-1,y_i,x,i)=\begin{cases}1, if\ y_i=[N],\ y_{i-1}=[V]\ and\ x_{i-1}=\text{"喝"}\\ 0,\ otherwise\end{cases}$$

$$s_l(y_i,x,i)=\begin{cases}1, if\ y_i=[V]\ and\ x_i=\text{"喝"}\\ 0,\ otherwise\end{cases}$$

计算特征函数在特征函数在各个位置i的求和，即$f_m(y,x)$

CRR的三个基本问题：

- 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$P(Y_i=y_i|x)$，$P(Y_{i-1}=y_{i-1},Y_i=y_i|x)$以及相应的数学期望的问题。采用前向-后向算法。
- 学习算法，学习权重系数，即式3.3中的λ,μ。采用极大似然估计和正则化的极大似然估计。
- 预测算法，给定条件随机场$P(Y|X)$，输入序列x和输出序列y，计算使得条件概率$P(Y_i=y_i|x)$最大的y^*，即对观测序列进行标注。采用维特比算法。

各方法细节见《统计学习方法(第2版)》P224~234

总结：

由上文可知，CRF已MaxEnt为基础，用于处理线性序列问题，广泛应用于中文分词、实体命名识别、词性标注等自然语言处理任务。

参考文献：

李航 .统计学习方法(第二版). 清华大学出版社

宗成庆.统计自然语言处理(第二版).清华大学出版社

周志华.机器学习.清华大学出版社