机器学习笔记–条件随机场

空修菜

1 条件随机场 (CRF)

1.1 CRF 的一般形式

1. 条件随机场（conditional random filed）与隐马尔可夫模型 HMM 一样, 一般用于标注问题.

2. HMM 模型是生成模型, CRF 模型是判别模型. 因为 CRF 要处理的问题是：根据给定的观测列, 返回一个输出的概率, 即计算的是条件概率 $P(Y \mid X)$.

3. 条件随机场是在已知观测变量序列的条件下, 标记序列发生的概率. 令 $X = \{X_1, X_2, \cdots, X_n\}$ 为观测变量序列, 令 $Y = \{Y_1, Y_2, \cdots, Y_n\}$ 为对应的标记变量序列. 条件随机场的目标是构建条件概率模型 $P(Y \mid X)$.

4. 标记随机变量序列 Y 的成员之间可能具有某种结构:

   (1) 在 NLP 的词性标注任务中, 观测数据为单词序列, 标记为对应的词性序列 (即动词、名词等词性的序列), 标记序列具有线性的序列结构;

   (2) 在 NLP 的语法分析任务中, 观测数据为单词序列, 标记序列是语法树 (主谓关系、动宾关系等), 标记序列具有树形结构.

5. 条件随机场处理的一般是链式条件随机场:

   - 在上图中, X 是观测序列, X_i 表示的是词, Y_i 是对应的词性.
   - $Y = Y_1, Y_2, \cdots, Y_n$ 是标记节点.

6. 给定观测变量序列 $X = \{X_1, X_2, \cdots, X_n\}$, 链式条件随机场主要包含两种关于标记变量的团（clique）:

   (1) 单个标记变量 Y_i 与 X 构成的团: $\{Y_i, X\}$, $1 \leq i \leq n$;

   (2) 相邻标记变量 Y_i, Y_{i+1} 与 X 构成的团: $\{Y_i, Y_{i+1}, X\}$, $1 \leq i \leq n-1$;

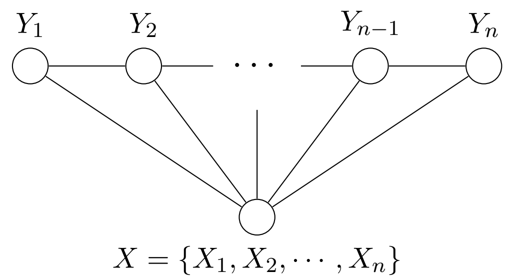

Figure 1.1: 线性链条件随机场

7. 条件随机场使用势函数和团来定义条件概率 $P(Y \mid X)$:

$$P(Y \mid X) = \frac{1}{Z} \exp\left(\sum_{j=1}^{K_1}\sum_{i=1}^{n-1} \lambda_j t_j(Y_i, Y_{i+1}, X, i) + \sum_{k=1}^{K_2}\sum_{i=1}^{n} \mu_k s_k(Y_i, X, i)\right)$$

(1) $t_j(Y_i, Y_{i+1}, X, i)$: 在已知观测序列的情况下, 两个相邻标记位置上的转移特征函数 (transition feature function):

- 它刻画了相邻标记变量之间的相邻关系, 以及观测序列 X 对它们的影响;
- i 是位置变量, 它是当前节点在标记序列中的位置. 比如, P 是代词, V 是代词, $i = 1$ 表示位置是句子的句首. 则:
  $t_j(Y_1 = P, Y_y = V, X, 1)$: 句首以介词开头, 紧接着的下一个词的词性是动词;
  $t_j(Y_1 = V, Y_y = P, X, 1)$: 句首以动词开头, 紧接着的下一个词的词性是介词;
  从经验来看, 代词位于动词前面的概率肯定比反过来高很多, 所以, $t_j(Y_1 = P, Y_y = V, X, 1)$ 的取值会更大, 一般取 1; $t_j(Y_1 = V, Y_y = P, X, 1)$ 的取值会更小, 一般取 0;
- K_1 是这个节点 (i) 的特征函数的总个数, $1 \leq j \leq K_1$, 每一个标记节点 i, 都有 K_1 个特征函数.

(2) $s_k(Y_i, X, i)$: 在已知观察序列情况下, 标记位置 i 上的状态特征函数 (status feature function).

- s_k 只依赖于当前的节点;
- K_2 是当前节点的状态特征函数的总个数;
- i 是当前标注序列的位置索引; N 表示名词, 则 $s_k(Y_1 = N, X, 1)$ 表示句首的词性是名词; $s_k(Y_1 = V, X, 1)$ 表示句首的词性是动词. 由经验可知, 名词在句首的概率比较大, 动词在句首的概率比较小;

(3) K_1, K_2 表示特征函数的总数, 一般没有一个固定数. 即可以规定有 3 个特征函数, 也可以规定有 10 个特征函数.

(4) 具体的例子:

$$t_j(Y_i, Y_{i+1}, X, i) = \begin{cases} 1 & if \quad Y_{i+1} = P,\ Y_i = V,\ X_i = \text{``}knock\text{''} \\ 0 & otherwise \end{cases}$$

$$s_k(Y_i, X, i) = \begin{cases} 1 & if \quad Y_i = V,\ X_i = \text{``}knock\text{''} \\ 0 & otherwise \end{cases}$$

- 转移特征函数刻画的是: 第 i 个观测值 X_i 为单词"knock"时, 相应的标记 Y_i, Y_{i+1} 很可能分别是动词 V 和介词 P;
- 状态特征函数刻画的是: 第 i 个观测值 X_i 是单词"knock"时, 标记 Y_i 很可能是动词 V.

1.2 CRF 的简化形式

1.2.1 原模型的简化形式

1. 转移特征函数与状态特征函数的统一形式.

   (1) 设有 K_1 个特征函数, K_2 个状态特征函数. $K = K_1 + K_2$, 记

   $$f_k(Y_i, Y_{i+1}, X, i) = \begin{cases} t_k(Y_i, Y_{i+1}, X, i) & k = 1, 2, \cdots, K_1 \\ s_l(Y_i, X, i) & k = K_1 + l,\ l = 1, 2, \cdots, K_2. \end{cases}$$

   注意上面的分段函数中, t_k 中的 i 的取值范围是: $1 \leq i < n$, 即不取 n, 而函数 s_l 可以取到 n.

   (2) 对转移特征与状态特征在每个 i 求和, 有

   $$f_k(Y, X) = \sum_{i=1}^{n} f_k(Y_i, Y_{i+1}, X, i), \quad k = 1, 2 \cdots, K$$

   其中, $Y = \{Y_1, Y_2, \ldots, Y_n\}$ 是标记变量序列, $X = \{X_1, X_2, \cdots, X_n\}$ 是观测变量序列.

2. 函数 f_k 的权重.

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

3. 利用以上两个简化记号, CRF 可以写为:

$$
\begin{aligned}
f(x) &= \frac{1}{Z}\exp\left(\sum_{j=1}^{K_1}\sum_{i=1}^{n-1}\lambda_j t_j(Y_i, Y_{i+1}, X, i) + \sum_{k=1}^{K_2}\sum_{i=1}^{n}\mu_k s_k(Y_i, X, i)\right) \\
&= \frac{1}{Z}\exp\left(\sum_{k=1}^{K_1}\lambda_k\sum_{i=1}^{n-1} t_k(Y_i, Y_{i+1}, X, i) + \sum_{l=1}^{K_2}\mu_l\sum_{i=1}^{n} s_l(Y_i, X, i)\right) \\
&= \frac{1}{Z}\exp\left(\sum_{k=1}^{K} w_k f_k(Y, X)\right),
\end{aligned}
$$

其中, $Z = \sum_Y \exp\left(\sum_{k=1}^{K} w_k f_k(Y, X)\right)$, $\sum_Y$ 表示对所有可能的标记序列进行求和.

4. 向量化表示 CRF 模型:

(1) 令权重向量为 w:

$$w = (w_1, w_2, \cdots, w_K)^T.$$

(2) 全局特征值向量 $F(Y, X)$ 为:

$$F(Y, X) = \big(f_1(Y, X), f_2(Y, X), \cdots, f_K(Y, X)\big)^T.$$

(3) 向量形式的 CRF 模型:

$$P(Y \mid X) = \frac{1}{Z}\exp\left(\sum_{k=1}^{K} w_k f_k(Y, X)\right) = \frac{1}{Z}\exp\left(w \cdot F(Y, X)\right),$$

其中,

$$Z = \sum_Y \exp\left(w \cdot F(Y, X)\right).$$

1.2.2 CRF 的矩阵形式

1. 设标记变量 $Y_i \in \Theta$, $\Theta = \{y_1, y_2, \cdots, y_m\}$, $1 \le i \le n$, 在词性标注问题中 Θ 可以看作是词性的集合, 即模型共有 m 个可能的词性. 对于观测变量序列 X 和标记变量序列的每个位置 $i = 0, 1, \cdots, n$, 定义一个 m 阶矩阵 $\mathbf{M}_i(X)$:

$$
\mathbf{M}_i(X) = \begin{pmatrix}
M_i(y_1, y_1 \mid X) & M_i(y_1, y_2 \mid X) & \dots & M_i(y_1, y_m \mid X) \\
M_i(y_2, y_1 \mid X) & M_i(y_2, y_2 \mid X) & \dots & M_i(y_2, y_m \mid X) \\
\vdots & \vdots & \ddots & \vdots \\
M_i(y_m, y_1 \mid X) & M_i(y_m, y_2 \mid X) & \dots & M_i(y_m, y_m \mid X)
\end{pmatrix}
$$

其中, $M_i(y_u, y_v \mid X) = M_i(Y_i = y_u, Y_{i+1} = y_v \mid X),\ 1 \le u, v \le m$.

$$M_i(Y_i = y_u, Y_{i+1} = y_v \mid X) = \exp\left(\sum_{k=1}^{K} w_k f_k(Y_i = y_u, Y_{i+1} = y_v, X, i),\right)$$

$\sum_{k=1}^{K} w_k f_k$ 的意思是：在位置 i, 所有转移特征函数值加上所有状态特征函数值.

2. 起点状态标记 $Y_0 = start$、终点状态标记 $Y_{n+1} = stop$.

   (1) $M_0(y_u, y_v \mid X)$ 表示将第 0 个位置标记为 y_u, 将第 1 个位置标记为 y_v, 则有

   $$\mathbf{M}_0(X) = \begin{pmatrix} M_0(start, y_1 \mid X) & M_0(start, y_2 \mid X) & \dots & M_0(start, y_m \mid X) \\ 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{pmatrix}$$

   (2) $M_n(y_u, y_v \mid X)$ 表示第 n 个位置标记为 y_u, 第 $n+1$ 个位置标记为 y_v, 则

   $$\mathbf{M}_n(X) = \begin{pmatrix} M_n(y_1, stop \mid X) & 0 & \dots & 0 \\ M_n(y_2, stop \mid X) & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ M_n(y_m, stop \mid X) & 0 & \dots & 0 \end{pmatrix}$$

3. 给定观测变量序列 X, 标记变量序列 $Y = \{Y_1, Y_2, \cdots, Y_n\}$ 可以按如下方式生成:

   (1) 开始位于起点状态 Y_0;

   (2) 然后从状态 Y_i 转移到状态 Y_{i+1}, $0 \le i \le n$;
   因此条件概率为:

   $$P(Y \mid X) = \frac{1}{Z} \prod_{i=0}^{n} \mathbf{M}_i(Y_i, Y_{i+1} \mid X),$$

   其中, Z 是以 $start$ 为起点, 以 $stop$ 为终点的所有标记路径的非规范化概率 $\prod_{i=0}^{n} \mathbf{M}_i(Y_i, Y_{i+1} \mid X)$ 之和:

   $$Z = \mathbf{M}_0(X)\mathbf{M}_1(X) \cdots \mathbf{M}_n(X),$$

   实际上, Z 取的是上面的矩阵乘积的第一行第一列的值.

1.3 CRF 的概率计算问题

1. 概率计算问题指的是：给定条件随机场模型 $P(Y \mid X)$, 给定输入序列 X, 给定输出序列 Y, 计算条件概率 $P(Y_i = y_i \mid X)$, $P(Y_{i-1} = y_{i-1}, Y_i = y_i \mid X)$, 以及相应的数学期望问题.

2. **前向向量**.

   (1) 定义. 对任意的 $i = 0, 1, \cdots, n+1$, 前向向量 $\alpha_i(X)$ 定义为:

   $$\alpha_0(Y_i \mid X) = \begin{cases} 1 & y = start \\ 0 & else. \end{cases}$$

   (2) 递归关系.

   $$\alpha_i^T(Y_i = y_u \mid X) = \alpha_{i-1}^T(Y_{i-1} = y_v \mid X)\mathbf{M}_i(Y_{i-1} = y_v, Y_i = y_u \mid X).$$

   (3) 含义：$\alpha_i^T(Y_i \mid X)$ 表示在位置 i 标记是 y_u, 且从 1 到 i 的非规范化概率. 因为 Y_i 有 m 个取值, 所以 $\alpha_i(X)$ 是 m 维的列向量.

3. **后向向量**.

   (1) 定义. 对任意的 $i = 0, 1, \cdots, n+1$, 后向向量 $\beta_i(X)$ 定义为:

   $$\beta_{n+1}(Y_{n+1} = y_{n+1} \mid X) = \begin{cases} 1 & y_{n+1} = start \\ 0 & else. \end{cases}$$

   (2) 递归关系.

   $$\beta_i(Y_i = y_u \mid X) = \mathbf{M}_{i+1}(Y_i = y_v, Y_{i+1} = y_u \mid X)\beta_{i+1}(Y_{i+1} = y_v \mid X).$$

   (3) 含义. $\beta_i^T(Y_i \mid X)$ 表示在位置 i 标记是 y_u, 且从 $i+1$ 到 n 的非规范化概率.

4. 条件概率 $P(Y_i = y_i \mid X)$ 的计算.

   $$P(Y_i = y_i \mid X) = \frac{\alpha_i^T(Y_i = y_i \mid X)\beta_i(Y_i = y_i \mid X)}{Z(X)},$$

   由 $\alpha_i(Y_i \mid X)$ 的定义, 它计算的是从 1 到 i, 且 $Y_i = y_i$ 时的非规范化概率; $\beta_i(Y_i \mid X)$ 计算的是从 $n+1$ 到 i, 且 $Y_i = y_i$ 时到非规范化概率. 即一个从前往后算, 另一个从后往前算, 两者在 i 处相遇, 此时, i 处的情形由两者共同决定.

   $$Z(X) = \alpha_n(X)^T\mathbf{1},$$

   $\mathbf{1}$ 是每一项都为 1 的 m 维向量.

5. 条件概率 $P(Y_{i-1}=y_{i-1}, Y_i=y_i \mid X)$ 的计算.

$$
\begin{aligned}
&P(Y_{i-1}=y_{i-1}, Y_i=y_i \mid X)\\
&=\frac{\alpha_{i-1}^T(Y_{i-1}=y_{i-1}\mid X)\mathbf{M}_i(Y_{i-1}=y_{i-1}, Y_i=y_i\mid X)\beta_i(Y_i=y_i\mid X)}{Z(X)}.
\end{aligned}
$$

6. 期望值的计算. 主要计算的是特征函数关于联合分布 $P(X,Y)$ 的期望 $\mathbb{E}_{P(X,Y)}[f_k]$ 以及关于条件分布 $P(Y\mid X)$ 的期望 $\mathbb{E}_{P(Y|X)}[f_k]$.

1.4 CRF 的学习算法

1.5 CRF 的预测算法

1. CRF 的预测问题指的是: 给定 CRF 模型 $P(Y\mid X)$ 和观测序列 X, 求使得条件概率最大的输出序列 (标记序列)Y^*. CRF 模型使用的是维特比算法.

2. 预测问题的最终形式:

$$
\begin{aligned}
Y^* &= \arg\max_Y P(Y\mid X) = \arg\max_Y \frac{\exp\big(w\cdot F(Y,X)\big)}{Z}\\
&= \arg\max_Y \exp\big(w\cdot F(Y,X)\big)\\
&= \arg\max_Y \big(w\cdot F(Y,X)\big)
\end{aligned}
$$

因此, 条件随机场的预测问题成为求非规范化概率的最优路径问题:

$$
\max_Y \big(w\cdot F(Y,X)\big),
$$

其中,

- $w=(w_1,w_2,\cdots,w_K)^T$;
- $F(Y,X)=\Big(f_1(Y,X), f_2(Y,X),\cdots, f_K(Y,X)\Big)^T$;
- $f_k(Y,X)=\sum_{i=1}^n f_k(Y_i,Y_{i+1},X,i),\ 1\le k\le K$.

3. 为了进一步简化模型符号, 定义局部特征向量 F_i:

$$
F_i(Y_i,Y_{i+1},X)=\big(f_1(Y_i,Y_{i+1},X,i), f_2(Y_i,Y_{i+1},X,i),\cdots, f_K(Y_i,Y_{i+1},X,i)\big).
$$

因此, 最终的问题可以写为

$$
Y^* = \arg\max_Y \big(w\cdot F(Y,X)\big) = \arg\max_Y \sum_{i=0}^n w\cdot F_i(Y_i,Y_{i+1},X).
$$

4. 维特比算法:

输入:模型的特征向量 $F(Y,X)$, 其中 Y 的取值集合为 $\{y_1,\cdots,y_m\}$. 权重向量 w, 观测序列 $X=\{X_1,\cdots,X_n\}$.

输出: 最优路径 $Y^*=\{y_1^*,y_2^*,\cdots,y_n^*\}$.

(1) 初始化:

$$\delta_1(j)=w\cdot F_0(Y_0=start,y_1=y_j,X),\ 1\le j\le m.$$

(2) 递推. 对 $1\le i\le n$:

$$\delta_i(l)=\max_{1\le j\le m}\{\delta_{i-1}(j)+w\cdot F_i(Y_i=y_j,Y_{i+1}=y_l,X)\},$$

$$\Psi_i(l)=\arg\max_{1\le j\le m}\{\delta_{i-1}(j)+w\cdot F_i(Y_i=y_j,Y_{i+1}=y_l,X)\},$$

$$1\le l\le m$$

(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:

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

最优路径为 $Y^*=\{y_1^*,y_2^*,\cdots,y_n^*\}$.