隐马尔可夫模型中最大似然估计的推导过程

夏庆荣，李正华

2015 年 11 月 9 日

1 符号定义

$\mathcal{D} = \{S^j, Y^j\}_{j=1}^N$：表示一个数据集，包含 N 个句子和对应的 N 个人工标注的词性序列。

$S^j = w_0^j w_1^j ... w_i^j ... x_{n_j}^j$：表示第 j 个句子，由 $n_j + 1$ 个词语组成。

$Y^j = y_0^j y_1^j ... y_i^j ... y_{n_j}^j$：表示第 j 个句子对应的词性序列。

为了简化数学表达，我们在每个句子开始和结束位置插入了伪词和伪词性，即：

$$\begin{aligned} w_0^j &= \textit{START} \\ y_0^j &= \textit{START} \\ w_{n_j}^j &= \textit{STOP} \\ y_{n_j}^j &= \textit{STOP} \end{aligned} \tag{1}$$

并且在学习和解码过程中，这些位置的词性都是固定的（不合法的发射概率和转移概率都为 0）。

$\mathcal{T}$：表示词性集合，即隐状态的所有可能取值，$y_i^j \in \mathcal{T}$。

$\mathcal{V}$：表示词表 (vocabulary)，即数据 $\mathcal{D}$ 所有词语的集合，$w_i^j \in \mathcal{V}$。

2 隐马尔可夫模型（hidden markov model, HMM）

根据贝叶斯公式，我们可以如下定义一个句子 $S = w_0...w_n$ 和对应词性序列 $Y = y_0...y_n$ 的联合概率（解释句子的生成过程）：

$$\begin{aligned} & p(S, Y) = p(w_0...w_n, y_0...y_n) \\ &= p(y_0...y_n) \times p(w_0...w_n | y_0...y_n) \\ &= \prod_{i=1}^{n} p(y_i | y_0...y_{i-1}) \times p(w_1...w_n | y_1...y_n) \end{aligned} \tag{2}$$

假设隐状态的转移过程满足一阶马尔科夫性质，即：

$$p(y_i | y_0...y_{i-1}) = p(y_i | y_{i-1})$$

同时假设词性序列 Y 生成句子 S 时，词语之间互相独立，并且词语 w_i 的生成过程只与当前词性 y_i 相关，即：

$$p(w_0...w_n | y_0...y_n) = \prod_{i=0}^{n} p(w_i | y_0...y_n) = \prod_{i=0}^{n} p(w_i | y_i) = \prod_{i=1}^{n} p(w_i | y_i)$$

这样，就可以有：

$$p(S,Y)=\prod_{i=1}^{n}[p(y_i|y_{i-1})\times p(w_i|y_i)] \tag{3}$$

进而，需要利用一个数据集 $\mathcal{D}$，估计语言模型所使用的所有参数：

$$\begin{aligned} q_{s,t} &\equiv p(t|s):\forall s\in\mathcal{T},\forall t\in\mathcal{T} \\ e_{t,w} &\equiv p(w|t):\forall t\in\mathcal{T},\forall w\in\mathcal{V} \end{aligned} \tag{4}$$

为了便于后续推导和理解，我们有时候会使用变量的形式，即 $q_{s,t}$ 和 $e_{t,w}$，分别表示转移概率和发射概率。可以看到，模型需要估计 $|\mathcal{T}|^2+|\mathcal{T}||V|$ 个参数。

根据如下公式确定参数通常称为最大似然估计 (maximum likehood estimation, MLE)：

$$\begin{aligned} p(t|s) &= \frac{\mathit{Count}(s,t)}{\mathit{Count}(s)} \\ p(w|t) &= \frac{\mathit{Count}(t,w)}{\mathit{Count}(t)} \end{aligned} \tag{5}$$

其中，$\mathit{Count}(s,t)$ 表示 st 这个 bigram（两个连续出现的词性）在数据集 D 中出现的次数；$\mathit{Count}(s)$ 表示 s 这个词性在数据集 $\mathcal{D}$ 中出现的次数；$\mathit{Count}(t,w)$ 表示 t,w 这个 bigram（一个词性及其相对应的词）在数据集中出现的次数；$\mathit{Count}(t)$ 表示 t 这个词性在数据集 $\mathcal{D}$ 中出现的次数。可以形式化表示如下：

$$\begin{aligned} \mathit{Count}(s,t) &= \sum_{j=1}^{N}\sum_{i=1}^{n_j}1[y_{i-1}^j=s\&\&y_i^j=t] \\ \mathit{Count}(w,t) &= \sum_{j=1}^{N}\sum_{i=1}^{n_j}1[y_i^j=t\&\&w_i^j=w] \end{aligned} \tag{6}$$

其中 $1[\mathit{condition}]$ 为指示函数（indicator function），如果 *condition* 为 true，则为 1，否则为 0。

这个文档的目的就是通过公式推导，说明最大似然估计的含义：根据公式 (6) 确定的参数，恰好让数据集 $\mathcal{D}$ 的 likehood 最大。

3 MLE 目标函数

数据集 $\mathcal{D}$ 的似然（likehood）定义如下。所谓 likehood，和概率应该是类似的，是说一个数据集存在的可能性。

$$\begin{aligned} L(\mathcal{D}) &= p(\mathcal{D}) \\ &= \prod_{j=1}^{N}p(S^j,Y^j) \\ &= \prod_{j=1}^{N}\prod_{i=1}^{n_j}p(y_i^j|y_{i-1}^j)p(w_i^j|y_i^j) \end{aligned} \tag{7}$$

对目标函数求对数在最优化和实际编程中是一种很常用的技巧。去对数可以让乘法变成加法，也可以让浮点数运算避免溢出。最重要的是，取对数不会影响原目标函数求极值。数据集 $\mathcal{D}$ 的对数似然函数（log-likehood）为：

$$
\begin{aligned}
LL(\mathcal{D}) &= \log L(\mathcal{D}) \\
&= \log \prod_{j=1}^{N} \prod_{i=1}^{n_j} p(y_i^j|y_{i-1}^j) p(w_i^j|y_i^j) \\
&= \sum_{j=1}^{N} \log \prod_{i=1}^{n_j} p(y_i^j|y_{i-1}^j) p(w_i^j|y_i^j) \\
&= \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left[\log p(y_i^j|y_{i-1}^j) + \log p(w_i^j|y_i^j)\right] \\
&= \sum_{j=1}^{N} \sum_{i=1}^{n_j} \left[\log q_{y_{i-1}^j, y_i^j} + \log e_{y_i^j, w_i^j}\right]
\end{aligned} \tag{8}
$$

对于每一个 $s \in \mathcal{T}, t \in \mathcal{T}, w \in \mathcal{V}$，二元模型都需要估计其对应参数 $q_{s,t}, e_{t,w}$。我们将所有参数聚在一起：

$$
\begin{aligned}
\Theta &= \{q_{s,t} : \forall s \in \mathcal{T}, t \in \mathcal{T}\} \\
\Lambda &= \{e_{t,w} : \forall t \in \mathcal{T}, w \in \mathcal{V}\}
\end{aligned} \tag{9}
$$

Θ 可以看成一个二维参数矩阵，包含 $|\mathcal{V}|^2$ 个参数；Λ 可以看成一个二维矩阵，包含 $|\mathcal{T}||\mathcal{V}|$ 个参数。

极大似然估计的形式化定义如下：

$$
(\Theta^{MLE}, \Lambda^{MLE}) = \arg\max_{\Theta,\Lambda} LL(\mathcal{D}) \tag{10}
$$

最大化目标函数最直接的方法是求偏导，然后求导数为 0 时的变量值。直接对公式求导无法求解。

公式 (10) 是一个无约束目标函数，但是我们知道每一个参数其实是一个条件概率，因此是应该有约束的。

$$
\begin{aligned}
&(\Theta^{MLE}, \Lambda^{MLE}) = \operatorname*{argmax}_{\Theta,\Lambda} LL(\mathcal{D}) \\
&s.t. \quad \sum_{t \in \mathcal{T}} q_{s,t} = 1 : \quad \forall s \in \mathcal{T} \\
&0 \le q_{s,t} \le 1 : \quad \forall s \in \mathcal{T}, \forall t \in \mathcal{T} \\
&\sum_{w \in \mathcal{V}} e_{t,w} = 1 : \quad \forall t \in \mathcal{T} \\
&0 \le e_{t,w} \le 1 : \quad \forall t \in \mathcal{T}, \forall w \in \mathcal{V}
\end{aligned} \tag{11}
$$

4 MLE 目标函数求解

求解有约束最优化问题的最简单方法是利用拉格朗日乘数法。对于每一个 $s \in \mathcal{T}$、$t \in \mathcal{T}$ 引入一个拉格朗日因子，即：

$$
\theta = \{\theta_s, \forall s \in \mathcal{T}\}, \lambda = \{\lambda_t, \forall t \in \mathcal{T}\} \tag{12}
$$

进而公式 (10) 的有约束问题可以转化为一个无约束问题（无需考虑只涉及一个变量自身的约束）：

$$
\begin{aligned}
& LL'(\mathcal{D};\Theta,\theta,\Lambda,\lambda) = LL(\mathcal{D}) + \sum_{s\in\mathcal{T}} \theta_s \times (1 - \sum_{t\in\mathcal{T}} q_{s,t}) + \sum_{t\in\mathcal{T}} \lambda_t \times (1 - \sum_{w\in\mathcal{V}} e_{t,w}) \\
& (\Theta^{MLE},\theta^{MLE},\Lambda^{MLE},\lambda^{MLE}) = \operatorname*{argmax}_{\Theta,\theta,\Lambda,\lambda} LL'(\mathcal{D};\Theta,\theta,\Lambda,\lambda)
\end{aligned}
\tag{13}
$$

接下来需要对公式针对 $q_{s,t}$ 和 $e_{t,w}$ 求偏导。但是公式 (8) 对 $LL(\mathcal{D})$ 的定义并没有 $q_{s,t}$ 和 $e_{t,w}$，而是一些具体的 $q_{y^j_{i-1},y^j_i}$ 和 $e_{y^j_i,w^j_i}$。因此，需要进一步处理：

$$
\begin{aligned}
LL(\mathcal{D}) &= \sum_{j=1}^{N}\sum_{i=1}^{n_j} \log q_{y^j_{i-1},y^j_i} + \sum_{j=1}^{N}\sum_{i=1}^{n_j} \log e_{y^j_i,w^j_i} \\
&= \sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}} \left\{ \sum_{j=1}^{N}\sum_{i=1}^{n_j} \log q_{y^j_{i-1},y^j_i} \times 1[y^j_{i-1} = s \&\& y^j_i = t] \right\} \\
&\quad + \sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{V}} \left\{ \sum_{j=1}^{N}\sum_{i=1}^{n_j} \log e_{y^j_i,w^j_i} \times 1[y^j_i = t \&\& w^j_i = w] \right\} \\
&= \sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}} \mathit{Count}(s,t) \log q_{s,t} + \sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{V}} \mathit{Count}(t,w) \log e_{t,w}
\end{aligned}
\tag{14}
$$

这样，我们可以对 $q_{s,t}$ 求偏导：

$$
\begin{aligned}
& \frac{\partial LL'(\mathcal{D};\Theta,\theta,\Lambda,\lambda)}{\partial q_{s,t}} \\
&= \frac{\partial \sum_{s\in\mathcal{T}}\sum_{t\in\mathcal{T}} \mathit{Count}(s,t)\log q_{s,t}}{\partial q_{s,t}} + \\
& \frac{\partial \sum_{t\in\mathcal{T}}\sum_{w\in\mathcal{V}} \mathit{Count}(t,w)\log e_{t,w}}{\partial q_{s,t}} + \\
& \frac{\partial \sum_{s\in\mathcal{T}} \theta_s \times (1 - \sum_{t\in\mathcal{V}} q_{s,t})}{\partial q_{s,t}} + \\
& \frac{\partial \sum_{t\in\mathcal{T}} \lambda_t \times (1 - \sum_{w\in\mathcal{V}} e_{t,w})}{\partial q_{s,t}} \\
&= \frac{\mathit{Count}(s,t)}{q_{s,t}} - \theta_s
\end{aligned}
\tag{15}
$$

令偏导为 0，可以得到：

$$
\begin{aligned}
\frac{\mathit{Count}(s,t)}{q_{s,t}} - \theta_s = 0 \\
q_{s,t} = \frac{\mathit{Count}(s,t)}{\theta_s}
\end{aligned}
\tag{16}
$$

代入约束

$$
\begin{aligned}
\sum_{t\in\mathcal{T}} q_{s,t} &= 1 \\
\sum_{t\in Y} \frac{\textit{Count}(s,t)}{\theta_s} &= 1 \\
\frac{\sum_{t\in\mathcal{T}} \textit{Count}(s,t)}{\theta_s} &= 1 \\
\frac{\textit{Count}(s)}{\theta_s} &= 1 \\
\theta_s &= \textit{Count}(s)
\end{aligned}
\tag{17}
$$

代入公式 (21)，得到：

$$
q_{s,t} = \frac{\textit{Count}(s,t)}{\textit{Count}(s)} \tag{18}
$$

接着，我们对 $e_{t,w}$ 求偏导。

$$
\begin{aligned}
&\frac{\partial LL'(\mathcal{D};\Theta,\theta,\Lambda,\lambda)}{\partial e_{t,w}} \\
=&\frac{\partial \sum_{s\in\mathcal{T}} \sum_{t\in\mathcal{T}} \textit{Count}(s,t) \log q_{s,t}}{\partial e_{t,w}} + \\
&\frac{\partial \sum_{t\in\mathcal{T}} \sum_{w\in\mathcal{V}} \textit{Count}(t,w) \log e_{t,w}}{\partial e_{t,w}} + \\
&\frac{\partial \sum_{s\in\mathcal{T}} \theta_s \times (1-\sum_{t\in\mathcal{T}} q_{s,t})}{\partial e_{t,w}} + \\
&\frac{\partial \sum_{t\in\mathcal{T}} \lambda_t \times (1-\sum_{w\in\mathcal{V}} e_{t,w})}{\partial e_{t,w}} \\
=&\frac{\textit{Count}(t,w)}{e_{t,w}} - \lambda_t
\end{aligned}
\tag{19}
$$

令偏导为 0，可以得到：

$$
\begin{aligned}
\frac{\textit{Count}(t,w)}{e_{t,w}} - \lambda_t = 0 \\
e_{t,w} = \frac{\textit{Count}(t,w)}{\lambda_t}
\end{aligned}
\tag{20}
$$

代入约束

$$
\begin{aligned}
\sum_{w \in \mathcal{V}} e_{t,w} &= 1 \\
\sum_{w \in \mathcal{V}} \frac{\textit{Count}(t,w)}{\lambda_t} &= 1 \\
\frac{\sum_{w \in \mathcal{V}} \textit{Count}(t,w)}{\lambda_t} &= 1 \\
\frac{\textit{Count}(t)}{\lambda_t} &= 1 \\
\lambda_t &= \textit{Count}(t)
\end{aligned}
\tag{21}
$$

代入公式 (26)，得到：

$$
e_{t,w} = \frac{\textit{Count}(t,w)}{\textit{Count}(t)} \tag{22}
$$

最终可以得到：

$$
\begin{aligned}
p(t|s) &= \frac{\textit{Count}(s,t)}{\textit{Count}(s)} \\
p(w|t) &= \frac{\textit{Count}(t,w)}{\textit{Count}(t)}
\end{aligned}
\tag{23}
$$