自然语言处理的最大熵模型

常宝宝
北京大学计算语言学研究所，100871

（一）

日常生活中，很多事情的发生表现出一定的随机性，试验的结果往往是不确定的，而且也不知道这个随机现象所服从的概率分布，所有的只有一些试验样本或样本特征，统计学常常关心的一个问题，在这种情况下如何对分布作出一个合理的推断？根据样本信息对某个未知分布作出推断的方法，最大熵的方法就是这样一个方法。

最大熵原理是在 1957 年由 E.T.Jaynes 提出的，其主要思想是，在只掌握关于未知分布的部分知识时，应该选取符合这些知识但熵值最大的概率分布。因为在这种情况下，符合已知知识的概率分布可能不止一个。我们知道，熵定义的实际上是一个随机变量的不确定性，熵最大的时侯，说明随机变量最不确定，换句话说，也就是随机变量最随机，对其行为做准确预测最困难。从这个意义上讲，那么最大熵原理的实质就是，在已知部分知识的前提下，关于未知分布最合理的推断就是符合已知知识最不确定或最随机的推断，这是我们可以作出的唯一不偏不倚的选择，任何其它的选择都意味着我们增加了其它的约束和假设，这些约束和假设根据我们掌握的信息无法作出。

看一个简单的例子：设 $a \in \{x, y\}$ 且 $b \in \{0, 1\}$，要推断概率分布 $p(a,b)$，唯一所知道的信息是 $p(x,0) + p(y,0) = 0.6$，即：

***p*(a,b)**	**0**	**1**	
x	?	?	
y	?	?	
	0.6		**1.0**

由于约束条件很少，满足条件的分布有无数多个，例如下面的分布就是满足已知条件的一个分布：

***p*(a,b)**	**0**	**1**	
x	0.5	0.1	
y	0.1	0.3	
	0.6		**1.0**

但按照最大熵原则，上述分布却不是一个好的分布，因为这个分布的熵不是满足条件的所有分布中熵最大的分布。按照最大熵的原则，应该选择的下面的分布：

***p*(a,b)**	**0**	**1**	
x	0.3	0.2	
y	0.3	0.2	
	0.6		**1.0**

因为，最大熵原则要求，合理的分布应该同时满足要求：

（1） $p^* = \arg\max_{p \in P} H(p) = \arg\max_{p \in P}[-\sum_{a \in \{x,y\}, b \in \{0,1\}} p(a,b) \log p(a,b)]$

（2） $p(x,0) + p(y,0) = 0.6$

（3） $p(x,0) + p(x,1) + p(y,0) + p(y,1) = 1$

上述例子比较简单，通过观察就可以得到熵值最大的概率分布，即使不能观察得到，也可以通过解析的方法得到。可是对于很多复杂的问题，往往不能用一个解析的办法获得。

（二）

自然语言处理中很多问题都可以归结为统计分类问题，很多机器学习方法在这里都能找到应用，在自然语言处理中，统计分类表现在要估计类 a 和某上下文 b 共现的概率 $P(a,b)$，不同的问题，类 a 和上下文 b 的内容和含义也不相同。在词性标注中是类的含义是词性标注集中的词类标记，而上下文指的是当前被处理的词前面一个词及词类，后面一个词及词类或前后若干个词和词类。通常上下文有时是词，有时是词类标记，有时是历史决策等等。大规模语料库中通常包含 a 和 b 的共现信息，但 b 在语料库中的出现常常是稀疏的，要对所有可能的(a,b)计算出可靠的 $P(a,b)$，语料库规模往往总是不够的。问题是要发现一个方法，利用这个方法在数据稀疏的条件下可靠的估计 $P(a,b)$。不同的方法可能采用不同的估计方法。

最大熵的原则：将已知事实作为制约条件，求得可使熵最大化的概率分布作为正确的概率分布。若用 A 表示所有类的集合，B 表示所有上下文的集合，那么正确的 p 应满足下面两条：

（1）可以使熵最大化的 p。

$$\hat{p} = \arg\max_{p} H(p)$$

这里 $x=(a,b)$，$a \in A$，$b \in B$，$\varepsilon = A \times B$

（2）p 要服从从样本数据中已知的统计证据。

现在的问题是已知知识如何表示，语料库中包含的各种知识应如何在最大熵模型中得到体现？在最大熵模型中，通常采用特征的办法来表示证据，特征可定义为如下的二值函数：

$$f : \varepsilon \to \{0,1\}$$

若有 k 个特征，那么特征 j 对 p 的制约可以表示为：

$$E_p f_j = E_{\tilde{p}} f_j \qquad (1)$$

其中 1≤j≤k，$E_p f_j$ 表示在概率分布为 p 时，特征 f_j 的期望值。$E_{\tilde{p}} f_j$ 表示特征 f_j 的样本期望值。所以有：

$$E_p f_j = \sum_{x \in \varepsilon} p(x) f_j(x)$$

$$E_{\tilde{p}} f_j = \sum_{x \in \varepsilon} \tilde{p}(x) f_j(x)$$

($\tilde{p}(x)$ 在这里表示事件 x 在样本数据中的概率)

公式(1)的含义是在概率分布 p 的情况下，特征的期望值应该和从样本数据中得到特征的样本期望值一致。

用 P 表示所有满足特征约束条件的分布，根据最大熵原则，就是要在 P 中选择一个能使熵取最大值的概率分布，这可以表示为：

$$P = \{p \mid E_p f_j = E_{\tilde{p}} f, 1 \le j \le k\}$$

$$p^* = \arg\max_{p \in P} H(p)$$

但满足上述条件的概率分布是一个什么样的分布呢？已经证明满足上述条件的概率分布 p*具有如下的形式：

$$p^*(x) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(x)}, \quad 0 \le \alpha_j \le \infty \qquad (2)$$

π是归一常数，α_j 是模型参数，每一个特征 f_j 对应一个α_j，α_j 可以被看作表示特征 f_j 相对重要程度的权重。

最大熵模型的优点是：在建模时，试验者只需要集中精力选择特征，而不需要花费精力考虑如何使用这些特征。而且可以很灵活地选择特征，使用各种不同类型的特征，特征容易更换。利用最大熵建模，一般也不需要做在其它方法建模中常常使用的独立性假设，参数平滑可以通过特征选择的方式加以考虑，无需专门使用常规平滑算法单独考虑，当然也不排除使用经典平滑算法进行平滑。每个特征对概率分布的贡献则由参数α_j决定，该参数可以通过一定的算法迭代训练得到。

（三）

令：

$P = \{p \mid E_p f_j = E_{\tilde{p}} f_j, 1 \le j \le k\}$

$Q = \{p \mid p(x) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(x)}, 0 \le j \le \infty\}$

则可以证明，（2）中的分布唯一且具有最大熵。

相对熵： p 和 q 是两个概率分布，二者的相对熵定义为：

$$D(p,q) = \sum_{x \in \varepsilon} p(x) \log \frac{p(x)}{q(x)}$$

相对熵在信息论中也叫交叉熵，Kullback 熵、鉴别信息等，其直观含义是对于随机变量 X 若开始认为概率分布是 $q(x)$，那么采用另外一个概率分布 $p(x)$，这种变化导致观察者所获得的信息量。

引理 1：p 和 q 是两个概率分布，则 $D(p, q) \geqslant 0$ 且 $D(p, q)=0$ 当且仅当 $p=q$。

引理 2（毕达哥拉斯性质）：若 $p \in P$，$q \in Q$，$p^* \in P \cap Q$，则：

$D(p, q) = D(p, p^*) + D(p^*, q)$

（证明略）

定理 1：若 $p^* \in P \cap Q$，则 $p^* = \arg\max_{p \in P} H(p)$，且 p^*是唯一的。

（证明略）

（四）

在最大熵模型中，参数α_j可通过 GIS（Generalized Iterative Scaling）算法进行，GIS 算法要求：

(1) $\forall x \in \varepsilon, \sum_{j=}^{k} f_j(x) = C$，$C$ 是常数。

(2) $\forall x \in \varepsilon \; \exists f_j, f_j(x) = 1$

若条件(1)不满足，则令：

$$C = \max_{x \in \varepsilon} \sum_{j=1}^{k} f_j(x)$$

并增加一个“校正”特征 f_l，$l=k+1$，且

$$\forall x \in \varepsilon, f_l(x) = C - \sum_{j=1}^{k} f_j(x)$$

注意，$0 \leqslant f_l(x) \leqslant C$，不象其它特征，$f_l(x)$的取值可能大于 1。

GIS 算法

$$\alpha_j^{(0)} = 1$$

$$\alpha_j^{(n+1)} = \alpha_j^{(n)} \left[\frac{\widetilde{E}f_j}{E^{(n)}f_j} \right]^{\frac{1}{C}}$$

这里：

$$E^{(n)}f_j = \sum_{x \in \varepsilon} p^{(n)}(x) f_j(x)$$

$$p^{(n)}(x) = \pi \prod_{j=1}^{l} (\alpha_j^{(n)})^{f_j(x)}$$

在 GIS 算法中，每循环一次，需要计算一次 $\widetilde{E}f_j$ 和 $E^{(n)}f_j$，其中 $\widetilde{E}f_j$ 不难计算，假定样本集合为：

$$S = \{(a_1,b_1), (a_2,b_2), \ldots, (a_N,b_N)\}$$

则：

$$\widetilde{E}f_j = \frac{1}{N} \sum_{i=1..N} f_j(a_i,b_i)$$

因为有太多可能的(a,b)，为了减少计算量，因而采用下面的公式近似计算 $E^{(n)}f_j$：

$$E^{(n)}f_j \approx \sum_{i=1}^{N} \widetilde{p}(b_i) \sum_{a \in A} p^{(n)}(a \mid b_i) f_j(a,b_i)$$

GIS 算法应在迭代足够次数时结束。

IIS 算法是用于训练最大熵模型的另外一个改进算法，训练时无需有上述条件(1)的限制。

（五）

在自然语言处理中，要做的统计推断常常是一个条件分布，在条件分布中熵的计算采用条件熵，此时最大熵模型为满足下列条件的模型：

$$p^* = \arg\max_{p \in P} H(p)$$

$$P = \{p \mid E_p f_j = E_{\tilde{p}} f_j, 1 \le j \le k\}$$

$$E_p f_j = \sum_{a,b} \widetilde{p}(b) p(a \mid b) f_j(a,b)$$

$$H(p) = -\sum_{a,b} \widetilde{p}(b) p(a \mid b) \log p(a \mid b)$$

此时最大熵模型应为：

$$p^*(a \mid b) = \frac{1}{Z(b)} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$$

$$Z(b) = \sum_{a} \prod_{j=1}^{k} \alpha_j^{f_j(a,b)}$$

（六）

特征选择是一个要解决的问题，对于样本数据，可以设计成千上万的特征，但并非所有特征都是可靠的，有些特征和样本数据的多少有关系，在样本数据少的情况下，计算出的样本期望和真实期望并不一致，选择哪些特征将是一个很关键的问题。这个问题要通过特征选择算法加以解决，假定所有特征的集合是 F，特征选择算法要从中选择一个活动特征集合 S，

活动特征集合要尽可能准确反映样本信息，只包括那些期望可以准确估计的特征。

为了求得 S，通常采用一个逐步增加特征的办法进行，每一次要增加哪个特征取决于样本数据。例如，当前的特征集合是 S，满足这些特征的模型是 $C(S)$，增加一个特征 f 意味着求得 $C(S)$的一个子集，该子集中的模型满足 $E_p f = E_{\tilde{p}} f$ 。新的模型集合可以定义为 $C(S \cup f)$。特征选择过程中，活动集合越来越大，而模型集合越来越小。

（七）

词性标注的任务是根据上下文 b_i 求当前词 w_i 的词性 t_i，可以看作是对 $P(t_i|b_i)$作出统计推断，对给定的词串

$$score(T) = \prod_{i=1..n} p(t_i \mid b_i)$$

$$T^* = \arg\max_T score(T)$$

$$b_i = (w_i, w_{i-1}, w_{i-2}, w_{i+1}, w_{i+2}, t_{i-1}, t_{i-2})$$

利用最大熵模型训练 $P(t_i|b_i)$

采用 beam search 计算最大的词性序列。

特征定义举例:

$$f_j(t, b_i) = \begin{cases} 1 & 若 t = \mathrm{DET} \wedge w_i = that \\ 0 & 其它 \end{cases}$$

$$f_k(t, b_k) = \begin{cases} 1 & 若 t = VBG \wedge suffix(w_i) = ing \\ 0 & 其它 \end{cases}$$

对上述定义的词性标注特征 $E_{\tilde{p}} f_j$ 即为(DET, that)在训练语料中出现频率除以语料中词的数量。

…

参考文献

Berger, A.L., Della Pietra, S.A., Della Pietra, V.J., (1996), A Maximum Entropy Approach to Natural Language Processing, Computational Linguistics, Volume 22, No. 1

Charniak, E., A Maximum-Entropy-Inspired Parser, …

Collins, M., (1999), Head-Driven Statistical Models for Natural Language Processing, University of Pennsylvania, Ph.D. Dissertation

Margerman, D.M., (1995), Statistical Decision-Tree Models for Parsing, In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics

Ratnaparkhi, A., (1996), A Maximum Entropy Part of Speech Tagger. In conference of Empirical Methods in Natural Language Processing, University of Pennsylvania

（ $p(a \mid H) = \frac{1}{Z(H)} e^{\lambda_1(a,H) f_1(a,H) + \lambda_2(a,H) f_2(a,H) + \ldots \lambda_m(a,H) f_m(a,H)}$ ）