Institute for Research in Cognitive Science

A Simple Introduction to Maximum Entropy Models for Natural Language Processing

Adwait Ratnaparkhi

University of Pennsylvania
3401 Walnut Street, Suite 400A
Philadelphia, PA 19104-6228

May 1997

Site of the NSF Science and Technology Center for
Research in Cognitive Science

IRCS Report 97--08

A Simple Introduction to Maximum Entropy Models for Natural Language Processing

Adwait Ratnaparkhi
Dept. of Computer and Information Science
University of Pennsylvania
adwait@unagi.cis.upenn.edu

May 13, 1997

Abstract

Many problems in natural language processing can be viewed as linguistic classification problems, in which linguistic contexts are used to predict linguistic classes. *Maximum entropy* models offer a clean way to combine diverse pieces of contextual evidence in order to estimate the probability of a certain linguistic class occurring with a certain linguistic context. This report demonstrates the use of a particular maximum entropy model on an example problem, and then proves some relevant mathematical facts about the model in a simple and accessible manner. This report also describes an existing procedure called *Generalized Iterative Scaling*, which estimates the parameters of this particular model. The goal of this report is to provide enough detail to re-implement the maximum entropy models described in [Ratnaparkhi, 1996, Reynar and Ratnaparkhi, 1997, Ratnaparkhi, 1997] and also to provide a simple explanation of the maximum entropy formalism.

1 Introduction

Many problems in natural language processing (NLP) can be re-formulated as statistical classification problems, in which the task is to estimate the probability of "class" a occurring with "context" b, or $p(a, b)$. Contexts in NLP tasks usually include words, and the exact context depends on the nature of the task; for some tasks, the context b may consist of just a single word, while for others, b may consist of several words and their associated syntactic labels. Large text corpora usually contain some information about the cooccurrence of a's and b's, but never enough to completely specify $p(a, b)$ for all possible (a, b) pairs, since the words in b are typically sparse. The problem is then to find a method for using the sparse evidence about the a's and b's to reliably estimate a probability model $p(a, b)$.

Consider the *Principle of Maximum Entropy* [Jaynes, 1957, Good, 1963], which states that the correct distribution $p(a, b)$ is that which maximizes entropy, or "uncertainty", subject to the constraints, which represent "evidence", i.e., the facts known to the experimenter. [Jaynes, 1957] discusses its advantages:

> ...in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have.

More explicitly, if $\mathcal{A}$ denotes the set of possible classes, and $\mathcal{B}$ denotes the set of possible contexts, p should maximize the entropy

$$H(p) = -\sum_{x \in \mathcal{E}} p(x) \log p(x)$$

where $x = (a, b)$, $a \in \mathcal{A}$, $b \in \mathcal{B}$, and $\mathcal{E} = \mathcal{A} \times \mathcal{B}$, and should remain consistent with the evidence, or "partial information". The representation of the evidence, discussed below, then determines the form of p.

2 Representing Evidence

One way to represent evidence is to encode useful facts as *features* and to impose constraints on the values of those feature expectations. A feature is a binary-valued function on events: $f_j : \mathcal{E} \to \{0, 1\}$. Given k features, the constraints have the form

$$E_p f_j = E_{\tilde{p}} f_j \tag{1}$$

where $1 \leq j \leq k$. $E_p f_j$ is the model p's expectation of f_j:

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

and is constrained to match the observed expectation, $E_{\tilde{p}} f_j$:

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

where $\tilde{p}$ is the observed probability of x in some training sample $\mathcal{S}$. Then, a model p is consistent with the observed evidence if and only if it meets the k constraints specified in (1). The Principle of Maximum Entropy recommends that we use p^*,

$$\begin{aligned} P &= \{p \mid E_p f_j = E_{\tilde{p}} f_j,\ j = \{1 \dots k\}\} \\ p^* &= \arg\max_{p \in P} H(p) \end{aligned}$$

$p(a,b)$	0	1	
x	?	?	
y	?	?	
total	.6		1.0

Table 1: Task is to find a probability distribution p under constraints $p(x,0)+p(x,1)=.6$, and $p(x,0)+p(x,1)+p(y,0)+p(y,1)=1$

	0	1	
x	.5	.1	
y	.1	.3	
total	.6		1.0

Table 2: One way to satisfy constraints

since it maximizes the entropy over the set of consistent models P. Section 5 shows that p^* *must* have a form equivalent to:

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

where π is a normalization constant and the α_j's are the model parameters. Each parameter α_j corresponds to exactly one feature f_j and can be viewed as a "weight" for that feature.

3 A Simple Example

The following example illustrates the use of maximum entropy on a very simple problem. Suppose the task is to estimate a probability distribution $p(a,b)$, where $a \in \{x,y\}$ and $b \in \{0,1\}$. Furthermore suppose that the only fact known about p is that $p(x,0)+p(y,0)=.6$. (The constraint that $\sum_{a,b} p(a,b)=1$ is implicit since p is a probability distribution.) Table 1 represents $p(a,b)$ as 4 cells labelled with "?", whose values must be consistent with the constraints. Clearly there are (infinitely) many consistent ways to fill in the cells of table 1; one such way is shown in table 2. However, the Principle of Maximum Entropy recommends the assignment in table 3, which is the most non-committal assignment of probabilities that meets the constraints on p.

Formally, under the maximum entropy framework, the fact

$$p(x,0)+p(y,0)=.6$$

is implemented as a constraint on the model p's expectation of a feature f:

$$E_p f = .6 \tag{3}$$

	0	1	
x	.3	.2	
y	.3	.2	
total	.6		1.0

Table 3: The most "uncertain" way to satisfy constraints

where

$$E_p f = \sum_{a \in \{x,y\}, b \in \{0,1\}} p(a,b) f(a,b)$$

and where f is defined as follows:

$$f(a,b) = \begin{cases} 1 & \text{if } b = 0 \\ 0 & \text{otherwise} \end{cases}$$

The observed expectation of f, or $E_{\tilde{p}} f$, is .6. The objective is then to maximize

$$H(p) = - \sum_{a \in \{x,y\}, b \in \{0,1\}} p(a,b) \log p(a,b)$$

subject to the constraint (3).

Assuming that features always map an event (a, b) to either 0 or 1, a constraint on a feature expectation is simply a constraint on the sum of certain cells in the table that represents the event space. While the above constrained maximum entropy problem can be solved trivially (by inspection), an iterative procedure is usually required for larger problems since multiple constraints may overlap in ways that prohibit a closed form solution.

Features typically express a cooccurrence relation between something in the linguistic context and a particular prediction. For example, [Ratnaparkhi, 1996] estimates a model $p(a, b)$ where a is a possible part-of-speech tag and b contains the word to be tagged (among other things). A useful feature might be

$$f_j(a,b) = \begin{cases} 1 & \text{if } a = \texttt{DETERMINER} \text{ and } \mathit{currentword}(b) = \text{“that”} \\ 0 & \text{otherwise} \end{cases}$$

The observed expectation $E_{\tilde{p}} f_j$ of this feature would then be the number of times we would expect to see the word "that" with the tag DETERMINER in the training sample, normalized over the number of training samples.

The advantage of the maximum entropy framework is that experimenters need only focus their efforts on deciding *what* features to use, and not on *how* to use them. The extent to which each feature f_j contributes towards $p(a, b)$, i.e., its "weight" α_j, is automatically determined by the Generalized Iterative Scaling algorithm. Furthermore, any kind of contextual feature can be used in the model, e.g., the model in [Ratnaparkhi, 1996] uses features that look at tag bigrams and word prefixes as well as single words.

Section 4 discusses preliminary definitions, section 5 discusses the maximum entropy property of the model of form (2), section 6 discusses its relation to maximum likelihood estimation, and section 7 describes the Generalized Iterative Scaling algorithm.

4 Preliminaries

Definitions 1 and 2 introduce relative entropy and some relevent notation. Lemmas 1 and 2 describe properties of the relative entropy measure.

Definition 1 (Relative Entropy, or Kullback-Liebler Distance). *The relative entropy D between two probability distributions p and q is given by:*

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

Definition 2.

$$\begin{aligned}
\mathcal{A} &= \textit{set of possible classes} \\
\mathcal{B} &= \textit{set of possible contexts} \\
\mathcal{E} &= \mathcal{A} \times \mathcal{B} \\
\mathcal{S} &= \textit{finite training sample of events} \\
\tilde{p}(x) &= \textit{observed probability of } x \textit{ in } \mathcal{S} \\
p(x) &= \textit{the model } p\textit{'s probability of } x \\
f_j &= \textit{A function of type } \mathcal{E} \rightarrow \{0,1\} \\
E_p f_j &= \sum_{x\in\mathcal{E}} p(x) f_j(x) \\
E_{\tilde{p}} f_j &= \sum_{x\in\mathcal{E}} \tilde{p}(x) f_j(x) \\
P &= \{p \mid E_p f_j = E_{\tilde{p}} f_j,\ j = \{1 \ldots k\}\} \\
Q &= \{p \mid p(x) = \pi \prod_{j=1}^{k} \alpha_j^{f_j(x)},\ 0 < \alpha_j < \infty\} \\
H(p) &= \sum_{x\in\mathcal{E}} p(x) \log p(x) \\
L(p) &= \sum_{x\in\mathcal{E}} \tilde{p}(x) \log p(x)
\end{aligned}$$

Here $\mathcal{E}$ is the event space, p always denotes a probability distribution defined on $\mathcal{E}$, P is the set of probability distributions consistent with the constraints (1), Q is the set of probability distributions of form (2), $H(p)$ is the entropy of p, and $L(p)$ is proportional to the log-likelihood of the sample $\mathcal{S}$ according to the distribution p.

Lemma 1. *For any two probability distributions p and q, $D(p,q) \geq 0$, and $D(p,q) = 0$ if and only if $p = q$.*

Proof: See [Cover and Thomas, 1991].

Lemma 2 (Pythagorean Property). *Given P and Q from Definition 2, if $p \in P$, $q \in Q$, and $p^* \in P \cap Q$, then*

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

This fact is discussed in [Csiszar, 1975] and more recently in [Della Pietra et al., 1995]. The term "Pythagorean" reflects the fact that this property is equivalent to the Pythagorean theorem in geometry if p,p^*,and q are the vertices of a right triangle and D is the squared distance function.

Proof. Note that for any $r, s \in P$, and $t \in Q$,

$$\begin{aligned}
\sum_{x \in \mathcal{E}} r(x) \log t(x) &= \\
\sum_x r(x)[\log \pi + \sum_j f_j(x) \log \alpha_j] &= \\
\log \pi[\sum_x r(x)] + [\sum_j \log \alpha_j \sum_x r(x) f_j(x)] &= \\
\log \pi[\sum_x s(x)] + [\sum_j \log \alpha_j \sum_x s(x) f_j(x)] &= \\
\sum_x s(x)[\log \pi + \sum_j f_j(x) \log \alpha_j] &= \\
&= \sum_x s(x) \log t(x)
\end{aligned}$$

Use the above substitution, and let $p \in P$, $q \in Q$, and $p^* \in P \cap Q$:

$$\begin{aligned}
D(p,p^*) + D(p^*,q) &= \\
\sum_x p(x) \log p(x) - \sum_x p(x) \log p^*(x) + \sum_x p^*(x) \log p^*(x) - \sum_x p^*(x) \log q(x) &= \\
\sum_x p(x) \log p(x) - \sum_x p(x) \log p^*(x) + \sum_x p(x) \log p^*(x) - \sum_x p(x) \log q(x) &= \\
\sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) &= D(p,q)
\end{aligned}$$

□

5 Maximum Entropy

Lemmas 1 and 2 derive the maximum entropy property of models of form (2) that satisfy the constraints (1):

Theorem 1. *If* $p^* \in P \cap Q$, *then* $p^* = \arg\max_{p \in P} H(p)$. *Furthermore,* p^* *is unique.*

Proof. Suppose $p \in P$ and $p^* \in P \cap Q$. Let $u \in Q$ be the uniform distribution so that $\forall x \in \mathcal{E} \quad u(x) = \frac{1}{|\mathcal{E}|}$.

- Show that $H(p) \leq H(p^*)$:

  By Lemma 2,

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

  and by Lemma 1,

  $$\begin{aligned} D(p,u) &\geq D(p^*,u) \\ -H(p) - \log\frac{1}{|\mathcal{E}|} &\geq -H(p^*) - \log\frac{1}{|\mathcal{E}|} \\ H(p) &\leq H(p^*) \end{aligned}$$

- Show p^* is unique:

  $$H(p) = H(p^*) \Longrightarrow D(p,u) = D(p^*,u) \Longrightarrow D(p,p^*) = 0 \Longrightarrow p = p^*$$

□

6 Maximum Likelihood

Secondly, models of form (2) that satisfy (1) have an alternate explanation under the maximum likelihood framework:

Theorem 2. *If* $p^* \in P \cap Q$, *then* $p^* = \arg\max_{q \in Q} L(q)$. *Furthermore* p^* *is unique.*

Proof. Let $\tilde{p}(x)$ be the observed distribution of x in the sample $\mathcal{S}$, $\forall x \in \mathcal{E}$. Clearly $\tilde{p} \in P$.

Suppose $q \in Q$ and $p^* \in P \cap Q$.

- Show that $L(q) \leq L(p^*)$:

  By Lemma 2,

  $$D(\tilde{p},q) = D(\tilde{p},p^*) + D(p^*,q)$$

  and by Lemma 1,

  $$\begin{aligned} D(\tilde{p},q) &\geq D(\tilde{p},p^*) \\ -H(\tilde{p}) - L(q) &\geq -H(\tilde{p}) - L(p^*) \\ L(q) &\leq L(p^*) \end{aligned}$$

- Show p^* is unique:

$$L(q) = L(p^*) \Longrightarrow D(\tilde{p}, q) = D(\tilde{p}, p^*) \Longrightarrow D(p^*, q) = 0 \Longrightarrow p^* = q$$

$\square$

Theorems 1 and 2 state that if $p^* \in P \cap Q$, then $p^* = \arg\max_{p \in P} H(p) = \arg\max_{q \in Q} L(q)$, and that p^* is unique. Thus p^* can be viewed under both the maximum entropy framework as well as the maximum likelihood framework. This duality is appealing, since p^*, as a maximum likelihood model, will fit the data as closely as possible, while as a maximum entropy model, will not assume facts beyond those in the constraints (1).

7 Parameter Estimation

Generalized Iterative Scaling [Darroch and Ratcliff, 1972], or GIS, is a procedure which finds the parameters $\{\alpha_1 \ldots \alpha_k\}$ of the unique distribution $p^* \in P \cap Q$.

The GIS procedure requires the constraint that

$$\forall x \in \mathcal{E} \ \sum_{j=1}^{k} f_j(x) = C$$

where C is some constant. If this is not the case, choose C to be

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

and add a "correction" feature f_l, where $l = k + 1$, such that

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

Note that unlike the existing features, $f_l(x)$ ranges from 0 to C, where C can be greater than 1.

Furthermore, the GIS procedure assumes that all events have at least one feature that is active,

$$\forall x \in \mathcal{E} \ \exists f_j \ f_j(x) = 1$$

Theorem 3. *The following procedure will converge to $p^* \in P \cap Q$*

$$\begin{aligned} \alpha_j^{(0)} &= 1 \\ \alpha_j^{(n+1)} &= \alpha_j^{(n)} [\frac{\tilde{E} f_j}{E^{(n)} f_j}]^{\frac{1}{C}} \end{aligned} \tag{4}$$

where

$$\begin{aligned} E^{(n)} f_j &= \sum_{x \in \mathcal{E}} p^{(n)}(x) f_j(x) \\ p^{(n)}(x) &= \pi \prod_{j=1}^{l} (\alpha_j^{(n)})^{f_j(x)} \end{aligned}$$

See [Darroch and Ratcliff, 1972] for a proof of convergence. [Darroch and Ratcliff, 1972] also show that the likelihood is non-decreasing, i.e., that $D(\tilde{p}, p^{(n+1)}) \leq D(\tilde{p}, p^{(n)})$, which implies that $L(p^{(n+1)}) \geq L(p^{(n)})$. See [Della Pietra et al., 1995] for a description and proof of *Improved Iterative Scaling*, which finds the parameters of p^* without the use of a "correction" feature. See [Csiszar, 1989] for a geometric interpretation of GIS.

7.1 Computation

Each iteration of the GIS procedure requires the quantities $E_{\tilde{p}} f_j$ and $E_p f_j$. The computation of $E_{\tilde{p}} f_j$ is straightforward given the training sample $\mathcal{S} = \{(a_1, b_1), \ldots, (a_N, b_N)\}$, since it is merely a normalized count of f_j:

$$E_{\tilde{p}} f_j = \sum_{i=1}^{N} \tilde{p}(a_i, b_i) f_j(a_i, b_i) = \frac{1}{N} \sum_{i=1}^{N} f_j(a_i, b_i)$$

where N is the number of event *tokens* (as opposed to *types*) in the sample $\mathcal{S}$.

However, the computation of the model's feature expectation,

$$E^{(n)} f_j = \sum_{a,b \in \mathcal{E}} p^{(n)}(a, b) f_j(a, b)$$

in a model with k (overlapping) features could be intractable since $\mathcal{E}$ could consist of 2^k distinguishable events. Therefore, we use the approximation originally described in [Lau et al., 1993]:

$$E^{(n)} f_j \approx \sum_{i=1}^{N} \tilde{p}(b_i) \sum_{a \in \mathcal{A}} p^{(n)}(a|b_i) f_j(a, b_i) \tag{5}$$

which only sums over the contexts in $\mathcal{S}$, and not $\mathcal{E}$, and makes the computation tractable.

The procedure should terminate after a fixed number of iterations (e.g., 100), or when the change in log-likelihood is negligible.

The running time of each iteration is dominated by the computation of (5) which is $O(NPA)$, where N is the training set size, P is the number of predictions, and A is the average number of features that are active for a given event (a, b).

8 Conclusion

This report presents the relevant mathematical properties of a maximum entropy model in a simple way, and contains enough information to reimplement the models described in [Ratnaparkhi, 1996, Reynar and Ratnaparkhi, 1997, Ratnaparkhi, 1997]. This model is convenient for natural language processing since it allows the unrestricted use of contextual features, and combines them in a principled way. Furthermore, its generality allows experimenters to re-use it for different problems, eliminating the need to develop highly customized problem-specific estimation methods.

References

[Cover and Thomas, 1991] Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

[Csiszar, 1975] Csiszar, I. (1975). I-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability*, 3(1):146–158.

[Csiszar, 1989] Csiszar, I. (1989). A Geometric Interpretation of Darroch and Ratcliff's Generalized Iterative Scaling. *The Annals of Statistics*, 17(3):1409–1413.

[Darroch and Ratcliff, 1972] Darroch, J. N. and Ratcliff, D. (1972). Generalized Iterative Scaling for Log-Linear Models. *The Annals of Mathematical Statistics*, 43(5):1470–1480.

[Della Pietra et al., 1995] Della Pietra, S., Della Pietra, V., and Lafferty, J. (1995). Inducing Features of Random Fields. Technical Report CMU-CS95-144, School of Computer Science, Carnegie-Mellon University.

[Good, 1963] Good, I. J. (1963). Maximum Entropy for Hypothesis Formulation, Especially for Multidimensional Contingency Tables. *The Annals of Mathematical Statistics*, 34:911–934.

[Jaynes, 1957] Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, 106:620–630.

[Lau et al., 1993] Lau, R., Rosenfeld, R., and Roukos, S. (1993). Adaptive Language Modeling Using The Maximum Entropy Principle. In *Proceedings of the Human Language Technology Workshop*, pages 108–113. ARPA.

[Ratnaparkhi, 1996] Ratnaparkhi, A. (1996). A Maximum Entropy Part of Speech Tagger. In *Conference on Empirical Methods in Natural Language Processing*, University of Pennsylvania.

[Ratnaparkhi, 1997] Ratnaparkhi, A. (1997). A Statistical Parser Based on Maximum Entropy Models. To appear in The Second Conference on Empirical Methods in Natural Language Processing.

[Reynar and Ratnaparkhi, 1997] Reynar, J. C. and Ratnaparkhi, A. (1997). A Maximum Entropy Approach to Identifying Sentence Boundaries. In *Fifth Conference on Applied Natural Language Processing*, pages 16–19, Washington D.C.