Bridging the Gap between Naive Bayes and Maximum Entropy Text Classification*

Alfons Juan[1], David Vilar[2] and Hermann Ney[2]

[1] DSIC/ITI, Univ. Politècnica de València, E-46022 València, Spain
[2] Lehrstuhl für Informatik 6, RWTH Aachen, D-52056 Aachen, Germany

Abstract. The *naive Bayes* and *maximum entropy* approaches to text classification are typically discussed as completely unrelated techniques. In this paper, however, we show that both approaches are simply two different ways of doing parameter estimation for a common log-linear model of class posteriors. In particular, we show how to map the solution given by maximum entropy into an optimal solution for naive Bayes according to the conditional maximum likelihood criterion.

1 Introduction

The *naive Bayes* and *maximum entropy* text classifiers are well-known techniques for text classification [1, 2]. Both techniques work with text documents represented as word counts. Also, both are log-linear decision rules in which an independent parameter is assigned to each class-word pair so as to measure their relative degree of association. Apparently, the only significant difference between them is the training criterion used for parameter estimation: conventional (joint) maximum likelihood for naive Bayes and *conditional* maximum likelihood for (the dual problem of) maximum entropy [2, 3]. This notable similarity, however, seems to have passed unnoticed for most researchers in text classification and, in fact, naive Bayes and maximum entropy are still discussed as unrelated methods.

In this paper, we provide a direct, bidirectional link between the naive Bayes and maximum entropy models for class posteriors. Using this link, maximum entropy can be interpreted as a way to train the naive Bayes model with conditional maximum likelihood. This is shown in Section 3, after a brief review of naive Bayes in the next section. Empirical results are reported in Section 4, and some concluding remarks are given in Section 5.

2 Naive Bayes Model

We denote the class variable by $c = 1, \ldots, C$, the word variable by $d = 1, \ldots, D$, and a document of length L by $d_1^L = d_1 d_2 \cdots d_L$. The joint probability of occurrence of c,

* Work supported by the EC (FEDER) and the Spanish "Ministerio de Educación y Ciencia" under grants TIN2006-15694-CO2-01 (iTransDoc research project) and PR-2005-0196 (fellowship from the "Secretaría de Estado de Universidades e Investigación").

Juan A., Vilar D. and Ney H. (2007).
Bridging the Gap between Naive Bayes and Maximum Entropy Text Classification.
In *Proceedings of the 7th International Workshop on Pattern Recognition in Information Systems*, pages 59-65
DOI: 10.5220/0002425700590065

L and d_1^L may be written as:

$$p(c, L, d_1^L) = p(c)\, p(L)\, p(d_1^L \mid c, L) \tag{1}$$

where we have assumed that document length does not depend on the class.

Given the class c and the document length L, the probability of occurrence of any particular document d_1^L can be greatly simplified by making the so-called *naive Bayes* or *independence assumption:* the probability of occurrence of a word d_l in d_1^L does not depend on its position l or other words $d_{l'}$, $l' \neq l$,

$$p(d_1^L \mid c, L) = \prod_{i=1}^{L} p(d_i \mid c) \tag{2}$$

Using the above assumptions, we may write the *posterior* probability of a document belonging to a class c as:

$$p(c \mid L, d_1^L) = \frac{p(c, L, d_1^L)}{\sum_{c'} p(c', L, d_1^L)} \tag{3}$$

$$= \frac{\vartheta(c) \prod_{d=1}^{D} \vartheta(d \mid c)^{x_d}}{\sum_{c'} \vartheta(c') \prod_{d=1}^{D} \vartheta(d \mid c')^{x_d}} \tag{4}$$

$$\stackrel{\triangle}{=} p_\theta(c \mid \boldsymbol{x}) \tag{5}$$

where x_d is the count of word d in d_1^L, $\boldsymbol{x} = (x_1, \ldots, x_D)^t$, and θ is the set of unknown parameters, which includes $\vartheta(c)$ for the class c prior and $\vartheta(d \mid c)$ for the probability of occurrence of word d in a document from class c. Clearly, these parameters must be non-negative and satisfy the normalisation constraints:

$$\sum_c \vartheta(c) = 1 \tag{6}$$

$$\sum_{d=1}^{D} \vartheta(d \mid c) = 1 \qquad (c = 1, \ldots, C) \tag{7}$$

The Bayes' decision rule associated with model (5) is a log-linear classifier:

$$\boldsymbol{x} \rightarrow c_\theta(\boldsymbol{x}) = \arg\max_c \; p_\theta(c \mid \boldsymbol{x}) \tag{8}$$

$$= \arg\max_c \left\{ \log \vartheta(c) + \sum_d x_d \log \vartheta(d \mid c) \right\} \tag{9}$$

3 Naive Bayes Training and Maximum Entropy

Naive Bayes training refers to the problem of deciding (a criterion and) a method to compute an appropriate estimate for θ from a given collection of N labelled training samples $(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)$. A standard training criterion is the *joint* log-likelihood function:

$$L(\theta) = \sum_n \log p_\theta(\boldsymbol{x}_n, c_n) \tag{10}$$

$$= \sum_c N_c \log \vartheta(c) + \sum_d N_{cd} \log \vartheta(d \mid c) \tag{11}$$

where N_c is the number of documents in class c and N_{cd} is the number of occurrences of word d in training data from class c. It is well-known that the global maximum of (10) under constraints (6)-(7) can be computed in closed-form:

$$\hat{\vartheta}(c) = \frac{N_c}{N} \tag{12}$$

and

$$\hat{\vartheta}(d \mid c) = \frac{N_{cd}}{\sum_{d'} N_{cd'}} \tag{13}$$

This computation is usually preceded by a preprocessing step in which documents are normalised in length so as to avoid parameter estimates being excessively influenced by long documents [4]. After training, this preprocessing step is no longer needed since the decision rule (8) is invariant to length normalisation. In what follows, we will assume that documents are normalised to unit length, i.e. $\sum_d x_d = 1$.

In this paper, we are interested in the *conditional log-likelihood* criterion:

$$CL(\theta) = \sum_n \log p_\theta(c_n \mid \boldsymbol{x}_n) \tag{14}$$

which is to be maximised under constraints (6)-(7). To this end, consider the conventional maximum entropy text classification model, as defined in [2]:

$$p_\Lambda(c \mid \boldsymbol{x}) = \frac{\exp\left[\sum_i \lambda_i f_i(\boldsymbol{x}, c)\right]}{\sum_{c'} \exp\left[\sum_i \lambda_i f_i(\boldsymbol{x}, c')\right]} \tag{15}$$

where the set $\Lambda = \{\lambda_i\}$ includes, for each class-word pair $i = (c', d')$, a (free) parameter $\lambda_i \in \mathbb{R}$ for its associated feature:

$$f_i(\boldsymbol{x}, c) = f_{c'd'}(\boldsymbol{x}, c) = \begin{cases} x_{d'} & \text{if } c' = c \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

Given an arbitrary value of the lambdas, $\tilde{\Lambda} = \{\tilde{\lambda}_i\}$, we have:

$$p_{\tilde{\Lambda}}(c|\boldsymbol{x}) = \frac{\exp\left[\sum_d \tilde{\lambda}_{cd}\, x_d\right]}{\sum_{c'} \exp\left[\sum_d \tilde{\lambda}_{c'd} x_d\right]} \tag{17}$$

$$= \frac{\prod_d \tilde{\alpha}_{cd}^{x_d}}{\sum_{c'} \prod_d \tilde{\alpha}_{c'd}^{x_d}} \qquad \text{with: } \tilde{\alpha}_{cd} \stackrel{\triangle}{=} \exp(\tilde{\lambda}_{cd}) \tag{18}$$

$$= \frac{\prod_d \tilde{\vartheta}(c,d)^{x_d}}{\sum_{c'} \prod_d \tilde{\vartheta}(c',d)^{x_d}} \qquad \tilde{\vartheta}(c,d) \stackrel{\triangle}{=} \frac{\tilde{\alpha}_{cd}}{\sum_{c'}\sum_{d'} \tilde{\alpha}_{c'd'}} \tag{19}$$

$$= \frac{\tilde{\vartheta}(c) \prod_d \left[\frac{\tilde{\vartheta}(c,d)}{\tilde{\vartheta}(c)}\right]^{x_d}}{\sum_{c'} \tilde{\vartheta}(c') \prod_d \left[\frac{\tilde{\vartheta}(c',d)}{\tilde{\vartheta}(c)}\right]^{x_d}} \qquad \tilde{\vartheta}(c) \stackrel{\triangle}{=} \sum_d \tilde{\vartheta}(c,d) \tag{20}$$

$$= p_{\tilde{\theta}}(c \mid \boldsymbol{x}) \qquad \tilde{\vartheta}(d \mid c) \stackrel{\triangle}{=} \frac{\tilde{\vartheta}(c,d)}{\tilde{\vartheta}(c)} \tag{21}$$

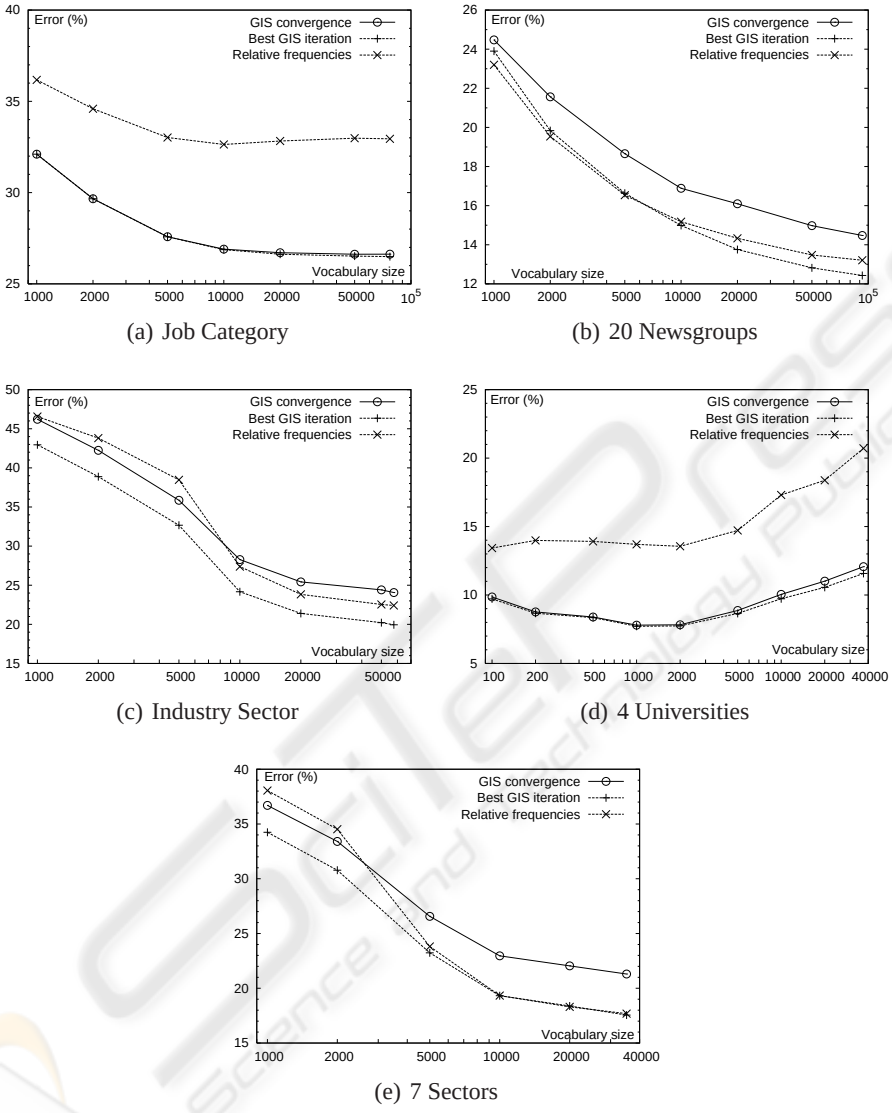
where, by definition, $\tilde{\theta}$ is non-negative and satisfy constraints (6)-(7).

Note that the definition given in (18) is a one-to-one mapping from $\tilde{\Lambda}$ to $\{\tilde{\alpha}_{cd}\}$. In contrast, that in (19) is a many-to-one mapping from $\{\tilde{\alpha}_{cd}\}$ to $\{\tilde{\vartheta}(c,d)\}$, though all possible $\{\tilde{\alpha}_{cd}\}$ mapping to the same $\{\tilde{\vartheta}(c,d)\}$ can be considered equivalent since they lead to the same class posterior distributions. Also note that $\{\tilde{\vartheta}(c,d)\}$ can be interpreted as the joint probability of occurrence of class c and word d. Thus, the mapping from $\{\tilde{\vartheta}(c,d)\}$ to $\tilde{\theta}$ defined in (20) and (21) is another one-to-one correspondence. All in all, the chain of equalities (17)-(21) and its associated definitions provide a direct, bidirectional link between the naive Bayes and maximum entropy models. In particular, to maximise (14) under constraints (6)-(7), it suffices to find a global optimum for the maximum entropy model and then map it to class priors and class-conditional word probabilities using the previous definitions.

4 Experiments

The experiments reported in this paper can be considered an extension of those reported in [2] and [5]. Our aim is to empirically compare conventional (joint) and conditional maximum likelihood training of the naive Bayes model. As in [5], we used the following datasets: *Job Category, 20 Newsgroups, Industry Sector, 7 Sectors and 4 Universities.* Table 1 contains some basic information on these datasets. For more details on them, please see [6], [7] and [5].

Preprocessing of the datasets was carried out with *rainbow* [8]. We used html skip for web pages, elimination of UU-encoded segments for newsgroup messages, and a special digit tagger for the *4 Universities* dataset [6]. We did not use stoplist removal, stemming or vocabulary pruning by occurrence count.



(a) Job Category (b) 20 Newsgroups

(c) Industry Sector (d) 4 Universities

(e) 7 Sectors

Fig. 1. Naive Bayes classification error rate as a function of the vocabulary size for the five datasets considered. Each plotted point is an error rate averaged over ten 80%-20% train-test splits. Each panel contains three curves: one corresponds to conventional parameter estimates (relative frequencies) and the other two refer to maximum entropy (conditional maximum likelihood) training using the GIS algorithm.

Table 1. Basic information on the datasets used in the experiments. (*Singletons* are words that occur once; *Class n-tons* refers to words that occur in n classes exactly).

	Job Category	20 Newsgroups	Industry Sector	4 Universities	7 Sectors
Type of documents	job titles & descriptions	newsgroup messages	web pages	web pages	web pages
Number of documents	131 643	19 974	9 629	4 199	4 573
Running words	11 221K	2 549K	1 834K	1 090K	864K
Average document length	85	128	191	260	189
Vocabulary size	84 212	102 752	64 551	41 763	39 375
Singletons (Vocab.%)	34.9	36.0	41.4	43.0	41.6
Classes	65	20	105	4	48
Class 1-tons (Vocab.%)	49.2	61.1	58.7	61.0	58.8
Class 2-tons (Vocab.%)	14.0	12.9	11.6	17.1	11.7

After preprocessing, ten random train-test splits were created from each dataset, with 20% of the documents held out for testing. Both, conventional and conditional maximum likelihood training of the naive Bayes model were compared in each split, using a training vocabulary comprising the top D most informative words in accordance to the *information gain criterion* [9] (D was varied from 100, 200, 500, 1000, ... up to full training vocabulary size). We used Laplace smoothing with $\epsilon = 10^{-5}$ for conventional training [5], and the GIS algorithm without smoothing for conditional maximum likelihood training through maximum entropy [10]. The results are shown in Figure 1. Each plotted point in this Figure is an error rate averaged over its corresponding ten data splits. Note that each plot contains one curve for the conventional training method and two curves for GIS training: one corresponds to the parameters obtained after the best iteration and the other to the parameters returned after GIS convergence. This "best iteration" curve may be interpreted as a (tight) lower bound to the error rate curve we could obtain by early stopping of the GIS to avoid overfitting.

From the results in Figure 1, we may say that conditional maximum likelihood training of the naive Bayes model provides similar to or better results than those of conventional training. In particular, they are significantly better in the Job Category and 4 Universities tasks, where it is also worth noting that maximum entropy does not suffer from overfitting (the best GIS iteration curve is almost identical to that after GIS convergence). However, in the 20 Newsgroups, Industry Sector and 7 Sectors tasks, the results are similar. Note that, in these tasks, the error curve for relative frequencies tends to lie in between the two curves for GIS, which are parallel and separated by a non-negligible offset (2% in 20 Newsgroups, and 4% in Industry Sector and 7 Sectors). Of course, this is a clear indication of overfitting that may be alleviated by early stopping of GIS and, as done for relative frequencies, by parameter smoothing. Another interesting conclusion we may draw from Figure 1 is that, with the sole exception of the 4 Universities task, the best results are obtained at full vocabulary size. This was previously observed in [5] for relative frequencies.

Summarising, the best test-set error rates obtained in the experiments are given in Table 2. These results match previous results usign the same techniques on the five

Table 2. Best test-set error rates for the five datasets considered.

	Parameter estimation		
Dataset	Smoothed relative frequencies	GIS after best iteration	GIS after convergence
Job category	32.6	26.3	26.4
20 Newsgroups	13.2	12.4	14.5
Industry-Sector	22.4	19.9	24.1
4 Universities	13.4	7.7	7.8
7 Sectors	17.7	17.6	21.3

datasets considered, though there are some minor differences due to different data preprocessing, experiment design or parameter smoothing [2, 5].

5 Conclusions

We have shown that the *naive Bayes* and *maximum entropy* text classifiers are closely related. More specifically, we have provided a direct, bidirectional link between the naive Bayes and maximum entropy models for class posteriors. Using this link, maximum entropy can be interpreted as a way to train the naive Bayes model with conditional maximum likelihood. We have extended previous empirical tests comparing these two training criteria. In summary, it may be said that conditional maximum likelihood training of the naive Bayes model provides similar to or better results than those of conventional training.

References

1. Lewis, D.: Naive Bayes at Forty: The Independence Assumption in Information Retrieval. In: Proc. of ECML-98. (1998) 4–15
2. Nigam, K., Lafferty, J., McCallum, A.: Using Maximum Entropy for Text Classification. In: Proc. of IJCAI-99 Workshop on Machine Learning for Information Filtering. (1999) 61–67
3. McCallum, A., Nigam, K.: A Comparison of Event Models for Naive Bayes Text Classification. In: Proc. of AAAI/ICML-98 Workshop on Learning for Text Categorization. (1998) 41–48
4. Juan, A., Ney, H.: Reversing and Smoothing the Multinomial Naive Bayes Text Classifier. In: Proc. of PRIS-02, Alacant (Spain) (2002) 200–212
5. Vilar, D., Ney, H., Juan, A., Vidal, E.: Effect of Feature Smoothing Methods in Text Classification Tasks. In: Proc. of PRIS-04, Porto (Portugal) (2004) 108–117
6. Ghani, R.: World Wide Knowledge Base (Web→KB) project. www-2.cs.cmu.edu/afs/cs.cmu.edu/project/theo-11/www/wwkb (2001)
7. Rennie, J.: Original 20 Newsgroups data set. (www.ai.mit.edu/∼jrennie) (2001)
8. McCallum, A.: Rainbow. (www.cs.umass.edu/∼mccallum/bow/rainbow) (1998)
9. Yang, Y., Pedersen, J.O.: A comparative study on feature selection in text categorization. In: Proc. of ICML-97. (1997) 412–420
10. Darroch, J., Ratcliff, D.: Generalized Iterative Scaling for Log-linear Models. Annals of Mathematical Statistics **43** (1972) 1470–1480