Proceedings of the 2003 Conference on Emprical Methods
in Natural Language Processing, pp. 137-144.

Evaluation and Extension of Maximum Entropy Models with Inequality Constraints

Jun'ichi Kazama†
kazama@is.s.u-tokyo.ac.jp
†Department of Computer Science
University of Tokyo
Hongo 7-3-1, Bunkyo-ku,
Tokyo 113-0033, Japan

Jun'ichi Tsujii†‡
tsujii@is.s.u-tokyo.ac.jp
‡CREST, JST
(Japan Science and Technology Corporation)
Honcho 4-1-8, Kawaguchi-shi,
Saitama 332-0012, Japan

Abstract

A maximum entropy (ME) model is usually estimated so that it conforms to equality constraints on feature expectations. However, the equality constraint is inappropriate for sparse and therefore unreliable features. This study explores an ME model with box-type inequality constraints, where the equality can be violated to reflect this unreliability. We evaluate the inequality ME model using text categorization datasets. We also propose an extension of the inequality ME model, which results in a natural integration with the Gaussian MAP estimation. Experimental results demonstrate the advantage of the inequality models and the proposed extension.

1 Introduction

The maximum entropy model (Berger et al., 1996; Pietra et al., 1997) has attained great popularity in the NLP field due to its power, robustness, and successful performance in various NLP tasks (Ratnaparkhi, 1996; Nigam et al., 1999; Borthwick, 1999).

In the ME estimation, an event is decomposed into *features*, which indicate the strength of certain aspects in the event, and the most uniform model among the models that satisfy:

$$E_{\tilde{p}}[f_i] = E_p[f_i], \quad (1)$$

for each feature. $E_{\tilde{p}}[f_i]$ represents the expectation of feature f_i in the training data (*empirical expectation*), and $E_p[f_i]$ is the expectation with respect to the model being estimated. A powerful and robust estimation is possible since the features can be as specific or general as required and does not need to be independent of each other, and since the most uniform model avoids overfitting the training data.

In spite of these advantages, the ME model still suffers from a lack of data as long as it imposes the equality constraint (1), since the empirical expectation calculated from the training data of limited size is inevitably unreliable. A careful treatment is required especially in NLP applications since the features are usually very sparse. In this study, text categorization is used as an example of such tasks with sparse features.

Previous work on NLP proposed several solutions for this unreliability such as the cut-off, which simply omits rare features, the MAP estimation with the Gaussian prior (Chen and Rosenfeld, 2000), the fuzzy maximum entropy model (Lau, 1994), and fat constraints (Khudanpur, 1995; Newman, 1977).

Currently, the Gaussian MAP estimation (combined with the cut-off) seems to be the most promising method from the empirical results. It succeeded in language modeling (Chen and Rosenfeld, 2000) and text categorization (Nigam et al., 1999). As described later, it relaxes constraints like $E_{\tilde{p}}[f_i] - E_p[f_i] = \frac{\lambda_i}{\sigma^2}$, where λ_i is the model's parameter.

This study follows this line, but explores the following box-type inequality constraints:

$$A_i \geq E_{\tilde{p}}[f_i] - E_p[f_i] \geq -B_i, \quad A_i, B_i > 0. \quad (2)$$

Here, the equality can be violated by the widths A_i and B_i. We refer to the ME model with the above inequality constraints as the *inequality ME* model.

This inequality constraint falls into a type of fat constraints, $a_i \leq E_p[f_i] \leq b_i$, as suggested by (Khudanpur, 1995). However, as noted in (Chen and Rosenfeld, 2000), this type of constraint has not yet been applied nor evaluated for NLPs.

The inequality ME model differs from the Gaussian MAP estimation in that its solution becomes sparse (i.e., many parameters become zero) as a result of optimization with inequality constraints. The features with a zero parameter can be removed from the model without changing its prediction behavior. Therefore, we can consider that the inequality ME model embeds feature selection in its estimation. Recently, the sparseness of the solution has been recognized as an important concept in constructing robust classifiers such as SVMs (Vapnik, 1995). We believe that the sparse solution improves the robustness of the ME model as well.

We also extend the inequality ME model so that the constraint widths can move using slack variables. If we penalize the slack variables by their 2-norm, we obtain a natural integration of the inequality ME model and the Gaussian MAP estimation. While it incorporates the quadratic stabilization of the parameters as in the Gaussian MAP estimation, the sparseness of the solution is preserved.

We evaluate the inequality ME models empirically, using two text categorization datasets. The results show that the inequality ME models outperform the cut-off and the Gaussian MAP estimation. Such high accuracies are achieved with a fairly small number of active features, indicating that the sparse solution can effectively enhance the performance. In addition, the 2-norm extended model is shown to be more robust in several situations.

2 The Maximum Entropy Model

The ME estimation of a conditional model $p(y|x)$ from the training examples $\{(x_i, y_i)\}$ is formulated as the following optimization problem.[1]

$$\begin{aligned} &\underset{p}{\text{maximize}}\ H(p) = \sum_x \tilde{p}(x) \sum_y p(y|x) \log p(y|x) \\ &\text{subject to}\ E_{\tilde{p}}[f_i] - E_p[f_i] = 0 \quad 1 \leq i \leq F. \end{aligned} \tag{3}$$

[1]To be precise, we have also the constraints $\sum_y p(y|x) - 1 = 0 \quad x \in \mathcal{X}$. Note that although we explain using a conditional model throughout the paper, the discussion can be applied easily to a joint model by considering the condition x is fixed.

The empirical expectations and model expectations in the equality constraints are defined as follows.

$$E_{\tilde{p}}[f_i] = \textstyle\sum_x \tilde{p}(x) \sum_y \tilde{p}(y|x) f_i(x, y), \tag{4}$$

$$E_p[f_i] = \textstyle\sum_x \tilde{p}(x) \sum_y p(y|x) f_i(x, y), \tag{5}$$

$$\tilde{p}(x) = c(x)/L, \ \ \tilde{p}(y|x) = c(x, y)/c(x), \tag{6}$$

where $c(\cdot)$ indicates the number of times $\cdot$ occurred in the training data, and L is the number of training examples.

By the Lagrange method, $p(y|x)$ is found to have the following parametric form:

$$p_\lambda(y|x) = \frac{1}{Z(x)} \exp(\sum_i \lambda_i f_i(x, y)), \tag{7}$$

where $Z(x) = \sum_y \exp(\sum_i \lambda_i f_i(x, y))$. The dual objective function becomes:

$$\begin{aligned} \mathcal{L}(\lambda) = &\textstyle\sum_x \tilde{p}(x) \sum_y \tilde{p}(y|x) \sum_i \lambda_i f_i(x, y) \\ &- \textstyle\sum_x \tilde{p}(x) \log \sum_y \exp(\sum_i \lambda_i f_i(x, y)). \end{aligned} \tag{8}$$

The ME estimation becomes the maximization of $\mathcal{L}(\lambda)$. And it is equivalent to the maximization of the log-likelihood: $LL(\lambda) = \log \prod_{x,y} p_\lambda(y|x)^{\tilde{p}(x,y)}$.

This optimization can be solved using algorithms such as the GIS algorithm (Darroch and Ratcliff, 1972) and the IIS algorithm (Pietra et al., 1997). In addition, gradient-based algorithms can be applied since the objective function is concave. Malouf (2002) compares several algorithms for the ME estimation including GIS, IIS, and the limited-memory variable metric (LMVM) method, which is a gradient-based method, and shows that the LMVM method requires much less time to converge for real NLP datasets. We also observed that the LMVM method converges very quickly for the text categorization datasets with an improvement in accuracy. Therefore, we use the LMVM method (and its variant for the inequality models) throughout the experiments. Thus, we only show the gradient when mentioning the training. The gradient of the objective function (8) is computed as:

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda_i} = E_{\tilde{p}}[f_i] - E_p[f_i]. \tag{9}$$

3 The Inequality ME Model

The maximum entropy model with the box-type inequality constraints (2) can be formulated as the fol-

lowing optimization problem:

$$\begin{aligned} \underset{p}{\text{maximize}} \ & \sum_x \tilde{p}(x) \sum_y p(y|x) \log p(y|x), \\ \text{subject to } & E_{\tilde{p}}[f_i] - E_p[f_i] - A_i \leq 0, \quad (10) \\ & E_p[f_i] - E_{\tilde{p}}[f_i] - B_i \leq 0. \quad (11) \end{aligned}$$

By using the Lagrange method for optimization problems with inequality constraints, the following parametric form is derived.

$$\begin{aligned} p_{\alpha,\beta}(y|x) &= \frac{1}{Z(x)} \exp(\sum_i (\alpha_i - \beta_i) f_i(x,y)), \\ \alpha_i &\geq 0, \quad \beta_i \geq 0, \quad (12) \end{aligned}$$

where parameters α_i and β_i are the Lagrange multipliers corresponding to constraints (10) and (11). The Karush-Kuhn-Tucker conditions state that, at the optimal point,

$$\begin{aligned} \alpha_i (E_{\tilde{p}}[f_i] - E_p[f_i] - A_i) &= 0, \\ \beta_i (E_p[f_i] - E_{\tilde{p}}[f_i] - B_i) &= 0. \end{aligned}$$

These conditions mean that the equality constraint is maximally violated when the parameter is non-zero, and if the violation is strictly within the widths, the parameter becomes zero. We call a feature *upper active* when $\alpha_i > 0$, and *lower active* when $\beta_i > 0$. When $\alpha_i - \beta_i \neq 0$, we call that feature *active*.[2] Inactive features can be removed from the model without changing its behavior. Since $A_i > 0$ and $B_i > 0$, any feature should not be upper active and lower active at the same time.[3]

The inequality constraints together with the constraints $\sum_y p(y|x) - 1 = 0$ define the feasible region in the original probability space, on which the entropy varies and can be maximized. The larger the widths, the more the feasible region is enlarged. Therefore, it can be implied that the possibility of a feature becoming inactive (the global maximal point is strictly within the feasible region with respect to that feature's constraints) increases if the corresponding widths become large.

[2]The term 'active' may be confusing since in the ME research, a feature is called active when $f_i(x,y) > 0$ for an event. However, we follow the terminology in the constrained optimization.

[3]This is only achieved with some tolerance in practice.

The solution for the inequality ME model would become sparse if the optimization determines many features as inactive with given widths. The relation between the widths and the sparseness of the solution is shown in the experiment.

The dual objective function becomes:

$$\begin{aligned} \mathcal{L}(\alpha,\beta) = & \textstyle\sum_x \tilde{p}(x) \sum_y \tilde{p}(y|x) \sum_i (\alpha_i - \beta_i) f_i(x,y) \\ & - \textstyle\sum_x \tilde{p}(x) \log \sum_y \exp(\sum_i (\alpha_i - \beta_i) f_i(x,y)) \\ & - \textstyle\sum_i \alpha_i A_i - \sum_i \beta_i B_i. \quad (13) \end{aligned}$$

Thus, the estimation is formulated as:

$$\underset{\alpha_i \geq 0, \beta_i \geq 0}{\text{maximize}} \ \mathcal{L}(\alpha,\beta).$$

Unlike the optimization in the standard maximum entropy estimation, we now have bound constraints on parameters which state that parameters must be non-negative. In addition, maximizing $\mathcal{L}(\alpha,\beta)$ is no longer equivalent to maximizing the log-likelihood $LL(\alpha,\beta)$. Instead, we maximize:

$$LL(\alpha,\beta) - \textstyle\sum_i \alpha_i A_i - \sum_i \beta_i B_i. \quad (14)$$

Although we can use many optimization algorithms to solve this dual problem since the objective function is still concave, a method that supports bounded parameters must be used. In this study, we use the BLMVM algorithm (Benson and Moré,), a variant of the limited-memory variable metric (LMVM) algorithm, which supports bound constraints.[4]

The gradient of the objective function is:

$$\begin{aligned} \frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \alpha_i} &= E_{\tilde{p}}[f_i] - E_p[f_i] - A_i, \\ \frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta_i} &= E_p[f_i] - E_{\tilde{p}}[f_i] - B_i. \quad (15) \end{aligned}$$

4 Soft Width Extension

In this section, we present an extension of the inequality ME model, which we call *soft width*. The soft width allows the widths to move as $A_i + \delta_i$ and $-B_i - \gamma_i$ using slack variables, but with some penalties in the objective function. This soft width extension is analogous to the soft margin extension of the SVMs, and in fact, the mathematical discussion is similar. If we penalize the slack variables

[4]Although we consider only the gradient-based method here as noted earlier, an extension of GIS or IIS to support bounded parameters would also be possible.

by their 2-norm, we obtain a natural combination of the inequality ME model and the Gaussian MAP estimation. We refer to this extension using 2-norm penalty as the *2-norm inequality ME* model. As the Gaussian MAP estimation has been shown to be successful in several tasks, it should be interesting empirically, as well as theoretically, to incorporate the Gaussian MAP estimation into the inequality model. We first review the Gaussian MAP estimation in the following, and then we describe our extension.

4.1 The Gaussian MAP estimation

In the Gaussian MAP ME estimation (Chen and Rosenfeld, 2000), the objective function is:

$$LL(\lambda) - \textstyle\sum_i (\frac{1}{2\sigma_i^2})\lambda_i^2, \tag{16}$$

which is derived as a consequence of maximizing the log-likelihood of the posterior probability, using a Gaussian distribution centered around zero with the variance σ_i^2 as a prior on parameters. The gradient becomes:

$$\frac{\partial \mathcal{L}(\lambda)}{\partial \lambda_i} = E_{\tilde{p}}[f_i] - E_p[f_i] - \frac{\lambda_i}{\sigma_i^2}. \tag{17}$$

At the optimal point, $E_{\tilde{p}}[f_i] - E_p[f_i] - \frac{\lambda_i}{\sigma_i^2} = 0$. Therefore, the Gaussian MAP estimation can also be considered as relaxing the equality constraints. The significant difference between the inequality ME model and the Gaussian MAP estimation is that the parameters are stabilized quadratically in the Gaussian MAP estimation (16), while they are stabilized linearly in the inequality ME model (14).

4.2 2-norm penalty extension

Our 2-norm extension to the inequality ME model is as follows.[5]

$$\begin{aligned} \underset{p,\delta,\gamma}{\text{maximize}}\ & H(p) - C_1 \textstyle\sum_i {\delta_i}^2 - C_2 \sum_i \gamma_i^2, \\ \text{subject to}\ & E_{\tilde{p}}[f_i] - E_p[f_i] - A_i \le \delta_i, \qquad (18) \\ & E_p[f_i] - E_{\tilde{p}}[f_i] - B_i \le \gamma_i, \qquad (19) \end{aligned}$$

where C_1 and C_2 is the penalty constants. The parametric form is identical to the inequality ME model (12). However, the dual objective function becomes:

$$LL(\alpha,\beta) - \textstyle\sum_i \left(\alpha_i A_i + \frac{\alpha_i^2}{4C_1}\right) - \sum_i \left(\beta_i B_i + \frac{\beta_i^2}{4C_2}\right).$$

Accordingly, the gradient becomes:

$$\begin{aligned} \frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \alpha_i} &= E_{\tilde{p}}[f_i] - E_p[f_i] - \left(A_i + \frac{\alpha_i}{2C_1}\right), \\ \frac{\partial \mathcal{L}(\alpha,\beta)}{\partial \beta_i} &= E_p[f_i] - E_{\tilde{p}}[f_i] - \left(B_i + \frac{\beta_i}{2C_2}\right). \end{aligned} \tag{20}$$

It can be seen that this model is a natural combination of the inequality ME model and the Gaussian MAP estimation. It is important to note that the solution sparseness is preserved in the above model.

5 Calculation of the Constraint Width

The widths, A_i and B_i, in the inequality constraints are desirably widened according to the unreliability of the feature (i.e., the unreliability of the calculated empirical expectation). In this paper, we examine two methods to determine the widths.

The first is to use a common width for all features fixed by the following formula.

$$A_i = B_i = W \times \tfrac{1}{L}, \tag{21}$$

where W is a constant, *width factor*, to control the widths. This method can only capture the global reliability of all the features. That is, only the reliability of the training examples as a whole can be captured. We call this method *single*.

The second, which we call *bayes*, is a method that determines the widths based on the Bayesian framework to differentiate between the features depending on their reliabilities.

For many NLP applications including text categorization, we use the following type of features.

$$f_{j,i}(x,y) = h_i(x) \text{ if } y = y_j, \ \ 0 \ \text{ otherwise.} \tag{22}$$

In this case, if we assume the approximation, $\tilde{p}(y|x) \approx \tilde{p}(y|h_i(x) > 0)$, the empirical expectation can be interpreted as follows.[6]

$$E_{\tilde{p}}[f_{j,i}] = \sum_{x:\ h_i(x)>0} \tilde{p}(x)\tilde{p}(y = y_j|h_i(x) > 0)h_i(x).$$

[5] It is also possible to impose 1-norm penalties in the objective function. It yields an optimization problem which is identical to the inequality ME model except that the parameters are upper-bounded as $0 \le \alpha_i \le C_1$ and $0 \le \beta_i \le C_2$. We will not investigate this 1-norm extension in this paper and leave it for future research.

[6] This is only for estimating the unreliability, and is not used to calculate the actual empirical expectations in the constraints.

Here, a source of unreliability is $\tilde{p}(y|h_i(x)>0)$. We consider $\tilde{p}(y|h_i(x)>0)$ as the parameter θ of the Bernoulli trials. That is, $p(y|h_i(x)>0) = \theta$ and $p(\bar{y}|h_i(x)>0) = 1-\theta$. Then, we estimate the posterior distribution of θ from the training examples by Bayesian estimation and utilize the variance of the distribution. With the uniform distribution as the prior, k times out of n trials give the posterior distribution: $p(\theta) = Be(1+k, 1+n-k)$, where $Be(\alpha, \beta)$ is the *beta* distribution. The variance is calculated as follows.

$$V[\theta] = \frac{(1+k)(1+n-k)}{(2+n)^2(n+3)}. \tag{23}$$

Letting $k = c(f_{j,i}(x,y)>0)$ and $n = c(h_i(x)>0)$, we obtain fine-grained variances narrowed according to $c(h_i(x)>0)$ instead of a single value, which just captures the global reliability. Assuming the independence of training examples, the variance of the empirical expectation becomes:

$$V\left[E_{\tilde{p}}[f_{j,i}]\right] = \left[\sum_{x:\, h_i(x)>0} \{\tilde{p}(x)h_i(x)\}^2\right] V[\theta_{j,i}].$$

Then, we calculate the widths as follows:

$$A_i = B_i = W \times \sqrt{V\left[E_{\tilde{p}}[f_{j,i}]\right]}. \tag{24}$$

6 Experiments

For the evaluation, we use the "Reuters-21578, Distribution 1.0" dataset and the "OHSUMED" dataset.

The Reuters dataset developed by David D. Lewis is a collection of labeled newswire articles.[7] We adopted "ModApte" split to split the collection, and we obtained $7,048$ documents for training, and $2,991$ documents for testing. We used 112 "TOPICS" that actually occurred in the training set as the target categories.

The OHSUMED dataset (Hersh et al., 1994) is a collection of clinical paper abstracts from the MEDLINE database. Each abstract is manually assigned MeSH terms. We simplified a MeSH term, like "$A/B/C \mapsto A$", and used the most frequent 100 simplified terms as the target categories. We extracted $9,947$ abstracts for training, and $9,948$ abstracts for testing from the file "ohsumed.91."

A documents is converted to a bag-of-words vector representation with TFIDF values, after the stop words are removed and all the words are downcased. Since the text categorization task requires that multiple categories are assigned if appropriate, we constructed a binary categorizer, $p_c(y \in \{+1, -1\}|d)$, for each category c. If the probability $p_c(+1|d)$ is greater than 0.5, the category is assigned. To construct a conditional maximum entropy model, we used the feature function of the form (22), where $h_i(d)$ returns the TFIDF value of the i-th word of the document vector.

We implemented the estimation algorithms as an extension of an ME estimation tool, Amis,[8] using the Toolkit for Advanced Optimization (TAO) (Benson et al., 2002), which provides the LMVM and the BLMVM optimization modules. For the inequality ME estimation, we added a hook that checks the KKT conditions after the normal convergence test.[9]

We compared the following models:

- ME models only with cut-off (*cut-off*),
- ME models with cut-off and the Gaussian MAP estimation (*gaussian*),
- Inequality ME models (*ineq*),
- Inequality ME models with 2-norm extension described in Section 4 (*2-norm*),[10]

For the inequality ME models, we compared the two methods to determine the widths, *single* and *bayes*, as described in Section 5. Although the Gaussian MAP estimation can use different σ_i for each feature, we used a common variance σ for *gaussian*. Thus, *gaussian* roughly corresponds to *single* in the way of dealing with the unreliability of features.

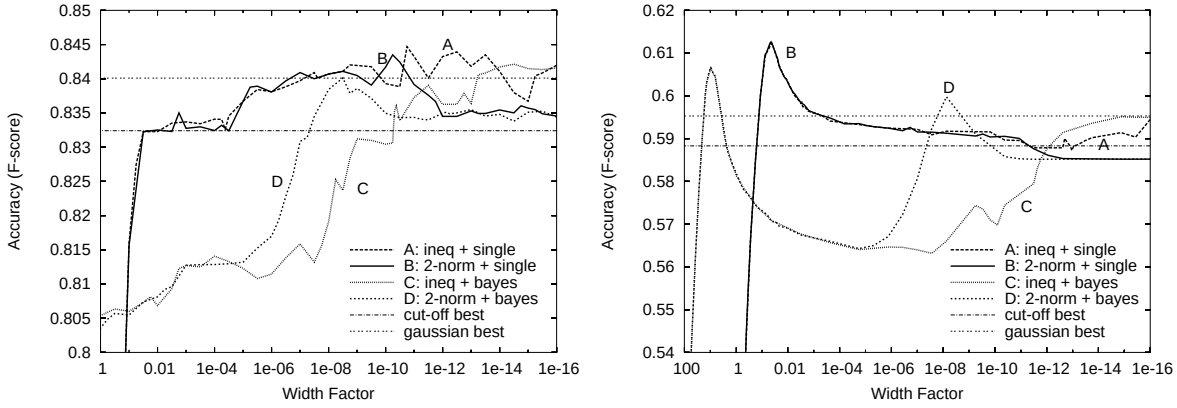
Note that, for inequality models, we started with all possible features and rely on their ability to remove unnecessary features automatically by solution sparseness. The average maximum number of features in a categorizer is $63,150.0$ for the Reuters dataset and $116,452.0$ for the OHSUMED dataset.

[7] Available from http://www.daviddlewis.com/resources/

[8] Developed by Yusuke Miyao so as to support various ME estimations such as the efficient estimation with complicated event structures (Miyao and Tsujii, 2002). Available at `http://www-tsujii.is.s.u-tokyo.ac.jp/~yusuke/amis`

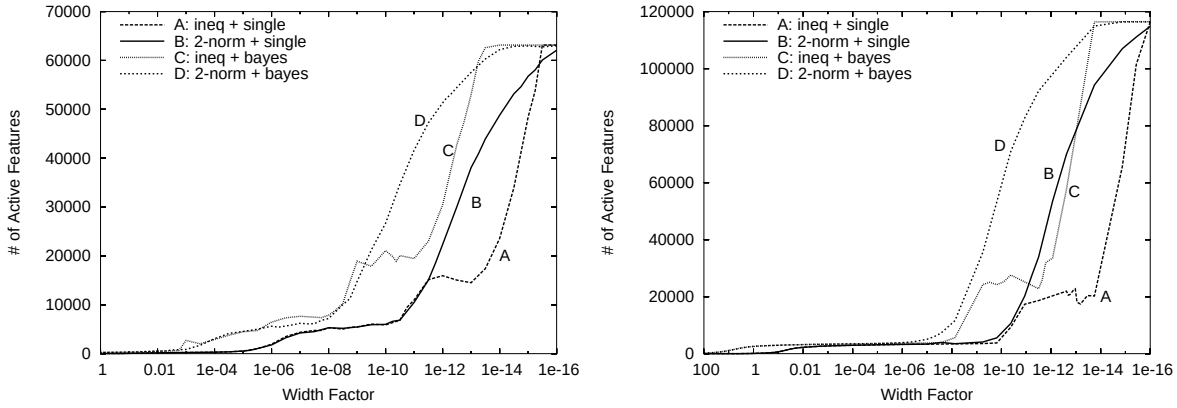
[9] The tolerance for the normal convergence test (relative improvement) and the KKT check is 10^{-4}. We stop the training if the KKT check has been failed many times and the ratio of the bad (upper and lower active) features among the active features is lower than 0.01.

[10] Here, we fix the penalty constants $C_1 = C_2 = 10^{16}$.



(a) Reuters (b) OHSUMED

Figure 1: Accuracies as a function of the width factor W for the development sets.



(a) Reuters (b) OHSUMED

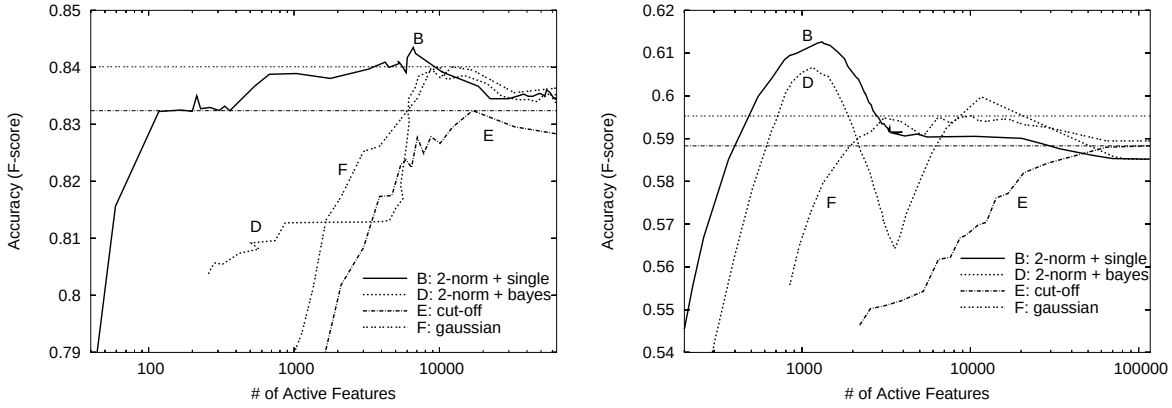
Figure 2: The average number of active features as a function of width factor W.

6.1 Results

We first found the best values for the control parameters of each model, W, σ, and the cut-off threshold, by using the development set. We show that the inequality models outperform the other methods in the development set. We then show that these values are valid for the evaluation set. We used the first half of the test set as the development set, and the second half as the evaluation set.

Figure 1 shows the accuracies of the inequality ME models for various width factors. The accuracies are presented by the "micro averaged" F-score. The horizontal lines show the highest accuracies of *cut-off* and *gaussian* models found by exhaustive search. For *cut-off*, we varied the cut-off threshold and found the best threshold. For *gaussian*, we varied σ with each cut-off threshold, and found the best σ and cut-off combination. We can see that the inequality models outperform the cut-off method and the Gaussian MAP estimation with an appropriate value for W in both datasets. Although the OHSUMED dataset seems harder than the Reuters dataset, the improvement in the OHSUMED dataset is greater than that in the Reuters dataset. This may be because the OHSUMED dataset is more sparse than the Reuters dataset. The 2-norm extension boosts the accuracies, especially for *bayes*, at the moderate Ws (i.e., with the moderate numbers of active features). However, we can not observe the apparent advantage of the 2-norm extension in terms of the highest accuracy here.

Figure 2 shows the average number of active features of each inequality ME model for various width factors. We can see that active features increase



(a) Reuters (b) OHSUMED

Figure 3: Accuracies as a function of the average number of active features for the development sets. For *gaussian*, the accuracy with the best σ found by exhaustive search is shown for each cut-off threshold.

when the widths become small as expected.

Figure 3 shows the accuracy of each model as a function of the number of active features. We can see that the inequality ME models achieve the highest accuracy with a fairly small number of active features, removing unnecessary features on their own. Besides, they consistently achieve much higher accuracies than the cut-off and the Gaussian MAP estimation with a small number of features.

Table 1 summarizes the above results including the best control parameters for the development set, and shows how well each method performs for the evaluation set with these parameters. We can see that the best parameters are valid for the evaluation sets, and the inequality ME models outperform the other methods in the evaluation set as well. This means that the inequality ME model is generally superior to the cut-off method and the Gaussian MAP estimation. At this point, the 2-norm extension shows the advantage of being robust, especially for the Reuters dataset. That is, the 2-norm models outperform the normal inequality models in the evaluation set. To see the reason for this, we show the average cross entropy of each inequality model as a function of the width factor in Figure 4. The average cross entropy was calculated as $-\frac{1}{C}\sum_c \frac{1}{L}\sum_i \log p_c(y_i|d_i)$, where C is the number of categories. The cross entropy of the 2-norm model is consistently more stable than that of the normal inequality model. Although there is no simple relation between the absolute accuracy and the cross entropy, this consistent difference can be one explanation for the advantage of the 2-norm extension. Besides, it is possible that the effect of 2-norm extension appears more clearly in the Reuters dataset because the robustness is more important in the Reuters dataset since the development set is rather small and easy to overfit.

Lastly, we could not observe the advantage of *bayes* method in these experiments. However, since our method is still in development, it is premature to conclude that the idea of using different widths according to its unreliability is not successful. It is possible that the uncertainty of $\tilde{p}(x)$, which were not concerned about, is needed to be modeled, or the Bernoulli trial assumption is inappropriate. Further investigation on these points must be done.

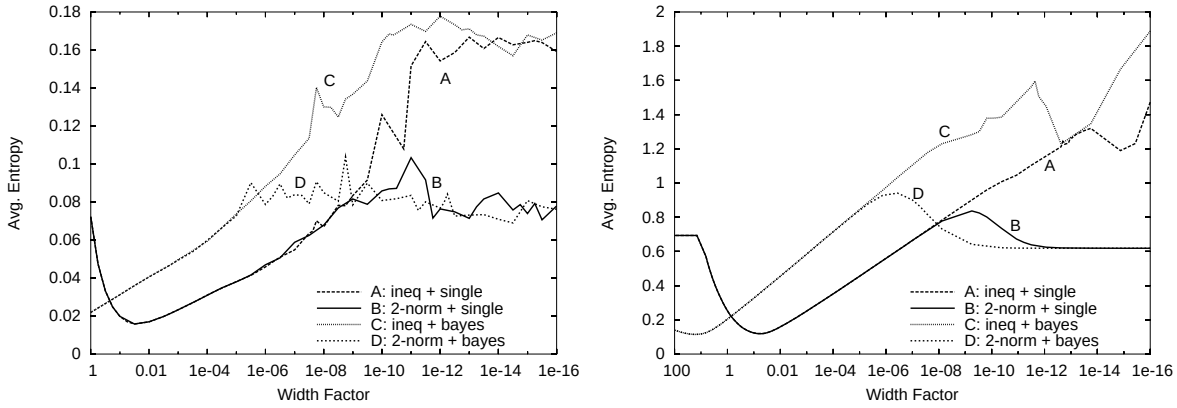
7 Conclusion and Future Work

We have shown that the inequality ME models outperform the cut-off method and the Gaussian MAP estimation, using the two text categorization datasets. Besides, the inequality ME models achieved high accuracies with a small number of features due to the sparseness of the solution. However, it is an open question how the inequality ME model differs from other sophisticated methods of feature selection based on other criteria.

Future work will investigate the details of the inequality model including the effect of the penalty constants of the 2-norm extension. Evaluations on other NLP tasks are also planned. In addition, we need to analyze the inequality ME model further to

Table 1: The summary of the experiments.

	Reuters				OHSUMED			
	best setting	# active feats	acc (dev)	acc (eval)	best setting	# active feats	acc (dev)	acc (eval)
cut-off	cthr $=2$	$16,961.9$	83.24	86.38	cthr $=0$	$116,452.0$	58.83	58.35
gaussian	cthr $=3$, $\sigma=4.22E3$	$12,326.6$	84.01	87.04	cthr $=8$, $\sigma=2.55E3$	$10,154.7$	59.53	59.08
ineq+single	$W=1.78E{-}11$	$9,479.9$	84.47	87.41	$W=4.22E{-}2$	$1,375.5$	61.23	61.10
2-norm+single	$W=5.62E{-}11$	$6,611.1$	84.35	87.59	$W=4.50E{-}2$	$1,316.5$	61.26	61.23
ineq+bayes	$W=3.16E{-}15$	$63,150.0$	84.21	87.37	$W=9.46$	$1,136.6$	60.65	60.31
2-norm+bayes	$W=3.16E{-}9$	$10,022.3$	84.01	87.57	$W=9.46$	$1,154.5$	60.67	60.32



(a) Reuters (b) OHSUMED

Figure 4: W vs. the average cross entropy for the development sets.

clarify the reasons for its success.

Acknowledgments We would like to thank Yusuke Miyao, Yoshimasa Tsuruoka, and the anonymous reviewers for many helpful comments.

References

S. J. Benson and J. J. Moré. A limited memory variable metric method for bound constraint minimization. Technical Report ANL/MCS-P909-0901, Argonne National Laboratory.

S. Benson, L. C. McInnes, J. J. Moré, and J. Sarich. 2002. TAO users manual. Technical Report ANL/MCS-TM-242-Revision 1.4, Argonne National Laboratory.

A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

A. Borthwick. 1999. A maximum entropy approach to named entity recognition. Ph.D. Thesis. New York University.

S. F. Chen and R. Rosenfeld. 2000. A survey of smoothing techniques for ME models. *IEEE Trans. on Speech and Audio Processing*, 8(1):37–50.

J. N. Darroch and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43:1470–1480.

W. Hersh, C. Buckley, T.J. Leone, and D. Hickam. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. In *Proc. of the 17th Annual ACM SIGIR Conference*, pages 192–201.

S. Khudanpur. 1995. A method of ME estimation with relaxed constraints. In *Johns Hopkins Univ. Language Modeling Workshop*, pages 1–17.

R. Lau. 1994. Adaptive statistical language modeling. A Master's Thesis. MIT.

R. Malouf. 2002. A comparison of algorithms for maximum entropy parameter estimation. In *Proc. of the sixth CoNLL*.

Y. Miyao and J. Tsujii. 2002. Maximum entropy estimation for feature forests. In *Proc. of HLT 2002*.

W. Newman. 1977. Extension to the ME method. In *IEEE Trans. on Information Theory*, volume IT-23, pages 89–93.

K. Nigam, J. Lafferty, and A. McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, pages 61–67.

S. Pietra, V. Pietra, and J. Lafferty. 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

A. Ratnaparkhi. 1996. A maximum entropy model for part-of-speech tagging. In *Proc. of the EMNLP*, pages 133–142.

V. Vapnik. 1995. *The Nature of Statistical Learning Theory*. Springer Verlag.