
Machine Learning, 60, 159–194, 2005

Maximum Entropy Models with Inequality Constraints: A Case Study on Text Categorization*

JUN'ICHI KAZAMA kazama@jaist.ac.jp
School of Information Science, Japan Advanced Institute of Science and Technology (JAIST), 1-1 Asahidai, Tatsunokuchi, Noumi, Ishikawa, 923-1292, Japan

JUN'ICHI TSUJII tsujii@is.s.u-tokyo.ac.jp
Department of Computer Science, Faculty of Information Science and Technology, University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 113-0033, Japan; CREST, JST (Japan Science and Technology Agency),, 4-1-8, Honcho, Kawaguchi, Saitama, 332-0012, Japan

Editors: Dan Roth and Pascale Fung

Abstract. Data sparseness or overfitting is a serious problem in natural language processing employing machine learning methods. This is still true even for the maximum entropy (ME) method, whose flexible modeling capability has alleviated data sparseness more successfully than the other probabilistic models in many NLP tasks. Although we usually estimate the model so that it completely satisfies the equality constraints on feature expectations with the ME method, complete satisfaction leads to undesirable overfitting, especially for sparse features, since the constraints derived from a limited amount of training data are always uncertain. To control overfitting in ME estimation, we propose the use of box-type inequality constraints, where equality can be violated up to certain predefined levels that reflect this uncertainty. The derived models, inequality ME models, in effect have regularized estimation with L_1 norm penalties of bounded parameters. Most importantly, this regularized estimation enables the model parameters to become sparse. This can be thought of as automatic feature selection, which is expected to improve generalization performance further. We evaluate the inequality ME models on text categorization datasets, and demonstrate their advantages over standard ME estimation, similarly motivated Gaussian MAP estimation of ME models, and support vector machines (SVMs), which are one of the state-of-the-art methods for text categorization.

Keywords: maximum entropy model, inequality constraint, regularization, feature selection, text categorization

1. Introduction

The maximum entropy (ME) method has gained a great deal of popularity in the NLP field since its introduction (Berger, Della Pietra, & Della Pietra 1996; Della Pietra, Della Pietra, & Lafferty, 1997). Applications include virtually all existing NLP tasks such as statistical machine translation (Berger, Della Pietra, & Della Pietra, 1996), part-of-speech (POS) tagging (Ratnaparkhi, 1996), text categorization (Nigam, John, & McCallum, 1999), named entity recognition (Borthwick, 1999), language modeling (Chen & Rosenfeld, 1999, 2000),

*This article is an elaborated version of (Kazama and Tsujii 2003). Several new experimental results have also been added.

and statistical parsing (Johnson et al., 1999; Johnson & Riezler, 2000). This popularity has been due to its powerful but robust modeling capabilities, simple and efficient learning algorithms, and empirical evidence that has demonstrated its advantages in various NLP tasks.

Robustness against data sparseness is the most advantageous property of ME models over traditional probabilistic models used for NLP such as naive Bayes models. To alleviate data sparseness, the ME method allows us to use *features* with various levels of specificity, which may be overlapping (e.g., uni-grams and bi-grams), and provides a means of estimation that can handle even these overlapping features based on the maximum entropy principle. However, the data sparseness problem cannot be solved completely even with the ME method. Thus, ways to avoid it, or control overfitting, are still important research topics for ME methods.

This paper also addresses the problem of data sparseness with the ME method. We especially focus on inappropriateness of equality constraints with the standard ME method, and propose the use of *inequality constraints*. The resulting ME model, which we call the *inequality ME* model, eventually becomes regularized and embeds feature selection in estimation, thus having good performance in generalization. In this study, we empirically demonstrate the advantages of inequality ME models through a text categorization task, which we consider suitable to evaluate the model's ability to alleviate data sparseness since it is a simple and standard task in evaluating machine learning techniques and the data sparseness problem, however, certainly exists.

1.1. Problem and existing solutions

With the ME method, an event is decomposed into features, each of which indicates the existence and strength of a certain aspect of the event. We estimate the model by viewing training examples through these features. Since features can be as specific or as general as required and do not need to be independent of one another, we can deal with the problem of data sparseness by using features of appropriate specificity (coverage). The maximum entropy principle states that estimation selects the model that has the highest entropy from models that satisfy, for all features, the equality constraint on the expected feature value:

$$E_{\tilde{p}}[f_i] - E_p[f_i] = 0. \tag{1}$$

$E_p[f_i]$ (*empirical expectation*) represents the expectation of feature f_i in the training data. Empirical expectation is the only information on the training data we can utilize for estimation. $E_p[\,f_i\,]$ (*model expectation*) is the expectation with respect to the model being estimated. When estimating, we maximize the model's entropy, while trying to satisfy the equality constraints. This estimation is based on the idea that the model that has the highest entropy of models that satisfy the constraints (i.e., the model that is the closest to the uniform distribution) is probably the best in terms of generalization performance (robustness).

However, maximum entropy estimation does not solve the data sparseness problem completely. As we will describe later, maximum entropy estimation is solved after being transformed to the maximum likelihood estimation (MLE) of the log-linear model. That is, the maximum entropy principle only restricts the models to the log-linear family and the overfitting problem in the MLE of log-linear family still exists. The problem can also be exposed from the viewpoint of constraints. Although the constraint in Eq. (1) is reasonable since, when we have a large (infinite) number of ideal training examples (without noise), the true model should satisfy the constraint, it is another cause of overfitting at the same time. In practice, empirical expectation inevitably contains uncertainty because it is calculated from limited training data, not from infinite amounts. Thus, complete satisfaction of equality is too strict a criterion. Complete satisfaction would be disastrous especially for sparse features. For example, if a feature occurs zero times in the training data, the probability of an event where the feature fires (has a non-negative value for the event) becomes zero, and the weights of that feature will be negative infinity. When predicting, the model will output zero probability for all input events where that feature fires. This is a kind of zero-frequency problem. It also cause numerical instability for ME estimation. To prevent this, we often omit zero-frequency features from the model beforehand. These concerns have motivated the need for feature selection to prevent the model from overfitting unreliable constraints.

Cut-off, which omits features that occur fewer than a predefined number of times, is a generalization of zero-frequency omission and one of the simplest ways of selecting features. Cut-off has been used in many applications with the ME method as a quick-fix. However, the problem with cut-off is that it omits features irrespective of whether low frequency is the result of data sparseness or it really indicates that they hold strong negative clues about prediction.

Many other feature selection methods have been proposed both for general settings (see, e.g., Yang & Pedersen, 1997, for a comparative study of these methods for text categorization) and for ME estimation (Berger, Della Pietra, & Della Pietra, 1996; Della Pietra, Della Pietra, & Lafferty, 1997; Shirai et al., 1998; McCallum, 2003; Zhou et al., 2003). They basically order and omit (or add) features, just by observing measures for the predictive power of features such as information gain, χ^2-test values, and gain in likelihood (Berger, Della Pietra, & Della Pietra, 1996; Della Pietra, Della Pietra, & Lafferty, 1997; McCallum, 2003; Zhou et al., 2003). The common problem with these methods is that the ordering is based on a heuristic criterion and ignores the fact that uncertainty is already contained in such measures. In addition, they are hardly able to consider the interaction between features because the features are selected sequentially. General, off-the-shelf, feature selection methods such as through information gain have another problem in that they ignore interaction with the model or the estimation algorithm. Since the goal is to maximize generalization performance with the model employed or estimation algorithm, the interaction between them should be taken into account. Although the feature selection methods for ME models (Berger, Della Pietra, & Della Pietra, 1996; Della Pietra, Della Pietra, & Lafferty, 1997; Shirai et al., 1998; McCallum 2003; Zhou et al., 2003) have interaction with the estimation algorithm, they abandon giving a complete account of feature interaction by resorting to approximation to avoid the expensive calculation of measures such as gain in likelihood.

Another major technique for controlling overfitting is regularized estimation, where the objective function is modified (regularized) so that overtraining can be avoided. The regularized approach is principled in that optimization is performed on the modified function exactly and no approximation is assumed. Several forms of regularization have been proposed for ME estimation such as maximum *a posteriori* (MAP) estimation with a Gaussian prior (Gaussian MAP estimation) (Chen & Rosenfeld, 1999; Johnson et al., 1999; Chen & Rosenfeld, 2000), the fuzzy maximum entropy model (Lau, 1994), fat constraints (Khudanpur, 1995; Newman, 1977), and, most recently, MAP estimation with an exponential prior (Goodman, 2003, 2004), which is very closely related to our method as we will describe in Section 6.

Empirically, Gaussian MAP estimation has been applied most successfully and shown to be useful in alleviating overfitting in various NLP tasks such as language modeling (Chen & Rosenfeld, 2000), text categorization (Nigam, John, & McCallum, 1999), part-of-speech tagging (Curran & Clark, 2003), shallow parsing (Sha & Pereira, 2003), and statistical parsing (Johnson et al., 1999; Johnson & Riezler, 2000). Gaussian MAP estimation alleviates overfitting by assuming a Gaussian prior on the parameter, which prevents excessively large or small weights. As a result, the objective function becomes a regularized log-likelihood, where the regularization term is the square of (weighted) L_2 norm of the parameters (i.e.,$-\sum_i \frac{1}{2\sigma_i^2}\lambda_i^2$). From the viewpoint of constraint satisfaction, Gaussian MAP estimation satisfies relaxed equality constraints:

$$E_{\tilde{p}}[f_i] - E_p[f_i] = \frac{\lambda_i}{\sigma_i^2}, \tag{2}$$

where λ_i is the model parameter and ${\sigma_i}^2$ is variance in the Gaussian prior. The ${\sigma_i}^2$ can be thought of as reflecting reliability in the expectation of feature f_i. Therefore, this approach considers unreliability of features more naturally than the feature selection methods we mentioned earlier. However, note that feature selection and regularization are not exactly different. They are similar in the sense that both approaches favor more simple (and therefore with high entropy) models. It is easy to see that fewer features yield higher entropy and reguralization approaches such as Gaussian MAP estimation (and our inequality ME estimation) find a model that has higher entropy than the model found by unregularized standard ME estimation. The difference lies in whether or not the uncertainity of the training data and feature interaction can be taken into account in a principled way. The similarity might be more easily understood by considering that selecting features corresponds to setting the regularization constants of Gaussian MAP estimation ($1/{\sigma_i}^2$) infinitely large, since the weights for such features should be zero.[1] The difference between the existing feature selection methods and the regularized estimation methods lies in whether or not the uncertainty of the training data and feature interaction can be taken into account in a principled way.

Although principled estimation in the regularization approach is appealing, omitting useless features explicitly also has an advantage in terms of model sizes, which affect the efficiency of processing and generalization performance, as many previous studies have demonstrated. Therefore, an interesting direction to take is clarifying whether feature selection and regularized estimation can be combined in a way where useless features

are automatically given zero weights by setting the regularized constants appropriately according to the uncertainty of the training data and whether such a method improves the performance over regularized estimation without feature selection ability. The inequality ME estimation, which we present in this paper, is one positive response to the above.

1.2. Our approach using inequality constraints

Our approach is regularization estimation, which is carefully designed to result in a kind of feature selection. Although the basic idea is to relax equality constraints as in Gaussian MAP estimation, we employ the following box-type inequality constraints.

$$-B_i \leq E_{\tilde{p}}[f_i] - E_p[f_i] \leq A_i$$
$$A_i > 0, B_i > 0. \tag{3}$$

Here, equality can be violated by predefined widths A_i and B_i. As the σ_i for Gaussian MAP estimation, A_i and B_i are suitably determined to reflect unreliability in feature f_i. These inequalities are the keys to enabling feature selection within the regularization approach. We refer to an ME model with these inequality constraints as an *inequality ME* model. It is an instance of the fat constraint model, and is a particular manifestation using the fat constraint, $a_i \leq E_p[f_i] \leq b_i$, as described by Khudanpur (1995). However, as noted by Chen and Rosenfeld (2000), this type of constraint has not yet been applied or evaluated since it was first suggested.

The parametric form of the inequality ME model becomes as simple as that of the standard model, except that each feature has two parameters α_i and β_i, where the first corresponds to the upper inequality constraint ($\leq$A$_i$) and the second corresponds to the lower inequality constraint ($-B_i \leq$). Training is also as simple as that for Gaussian MAP estimation. The objective function to estimate the inequality ME model also becomes a regularized log-likelihood, where the regularized term is the (weighed) L_1 norm of the parameters (i.e., $-\sum_i A_i\ \alpha_i - \sum_i B_i\ \beta_i$). However, the parameters are bounded, i.e., $\alpha_i, \beta_i \geq 0$ and algorithms that support such bounded parameters are required. The extra computational cost is acceptable, although there is actually extra cost since the bounded optimization will be more difficult and the number of parameters are doubled. As we will describe later, this model corresponds to using an exponential prior in MAP estimation[1] (Goodman, 2003, 2004).

The inequality ME model differs from Gaussian MAP estimation in that, as a result of using inequality constraints, the solution becomes sparse (i.e., many parameters become zero) by setting constraint widths A_i and B_i appropriately. As we demonstrate in the experiments, broader widths give a sparser solution. Features with a zero parameter could be removed from the model without changing its predictive behavior. Thus, the inequality ME model can be considered to embed feature selection in its estimation. Note that this feature selection is completely different from Gaussian MAP estimation with an infinitely large regularization constant (i.e., zero variance), which we mentioned earlier as a fictional scenario. This inequality ME estimation gives an exactly zero weight event if A_i and B_i

are not infinite unless the equality violation is maximal (see Section 3 for details). That is, inequality ME estimation favors zero weights much more than Gaussian MAP estimation.

This embedded feature selection was inspired by the derivation of support vector machines (SVMs) (Vapnik, 1995), where sparse solutions (called support vectors) are the result of inequality constraints in the problem definitions.[2] A sparse solution is recognized as an important concept in explaining state-of-the-art robustness in SVMs. We believe that the sparse solution also improves the robustness of the ME model. In fact, we demonstrate through experiments that the inequality ME model outperforms the standard ME model.

This paper also presents an extension of the inequality ME model, where constraint widths can move by using slack variables. If we penalize slack variables by the square of their L_2 norm (2-norm extension), we obtain a natural integration of inequality ME estimation and Gaussian MAP estimation. While it adds quadratic stabilization of parameters as in Gaussian MAP estimation, the sparseness of the solution is preserved. We also show that we obtain a inequality ME model where the parameters are bounded from above as well as from below when we penalize the slack variables by their L_1 norm (1-norm extension).

We evaluated inequality ME models empirically, using two text categorization datasets. The comparison in this study mainly focuses on how inequality ME estimation, which embeds feature selection, outperforms Gaussian MAP estimation, which does not. Thus, comparisons with other feature selection methods have been omitted. The experimental results revealed that the inequality ME models outperformed standard ME estimation and Gaussian MAP estimation. We also found that such high accuracies can be achieved with a fairly small number of active features, indicating that a sparse solution can effectively enhance performance. The 2-norm extended model was more robust in several situations. We also compared inequality ME models with SVMs, which have recently been thought of as state-of-the-art in many respects for NLP applications. We found that, at least in our experiments on text categorization, inequality ME models outperformed SVMs with a polynomial kernel and an RBF kernel, while the SVMs achieved higher accuracies than standard ME estimation or Gaussian MAP estimation.

The rest of the paper is organized as follows. Section 2 describes the maximum entropy model and its extension using Gaussian MAP estimation. Then, we describe the inequality ME model in Section 3. Extensions to the inequality ME method are described in Section 5. Section 4 discusses methods of determining the width of inequality constraints. Experiments on text categorization are described in Section 7. We conclude the paper with a discussion and outline plans for future work.

2. Preliminaries

In this section, we describe a conditional ME model (Berger, Della Pietra, & Della Pietra, 1996) and its extension using MAP estimation with a Gaussian prior (Chen & Rosenfeld, 1999, 2000; Johnson et al., 1999, 2000) to make the later explanation of inequality models easier and more self-contained. Although our explanation uses a conditional model throughout the paper, discussions such as the derivation of inequality ME models can easily be applied to the case of a joint model.

2.1. *Maximum Entropy Model*

Assume that our aim is to estimate the relation between x (input) and y (output) as a probabilistic model $p(y \mid x)$ from the training examples $\{(x_k, y_k)\}_{k=1}^{L}$. For text categorization, x represents a document and y represents a category label (or ± 1 labels that indicates whether the document belong to a certain category or not). In ME estimation, we define a set of feature functions $F = \{f_i(x, y)\}$ to model an event $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Each $f_i(x, y)$ indicates a certain aspect of the event (x,y) and its strength through a non-negative value (i.e., $f_i(x,y) \geq 0$). ME estimation finds the model with the highest entropy from models that satisfy the equality between empirical expectation and model expectation, which are defined as follows.

$$E_{\tilde{p}}[f_i] = \sum_x \tilde{p}(x) \sum_y \tilde{p}(y \mid x) f_i(x, y) \quad \text{(empirical expectation) and} \tag{4}$$

$$E_p[f_i] = \sum_x \tilde{p}(x) \sum_y p(y \mid x) f_i(x, y) \quad \text{(model expectation).} \tag{5}$$

Empirical expectation is the expectation of the value of a feature in the training data, and model expectation is expectation by the model being estimated. $\tilde{p}(x)$ and $\tilde{p}(y \mid x)$ are called empirical distributions and calculated as follows

$$\tilde{p}(x) = c(x)/L,$$
$$\tilde{p}(y \mid x) = c(x, y)/c(x).$$

L is the number of training examples, and $c(x)$ indicates the number of times x occurs in the training data.

Then, ME estimation is formulated as the following optimization problem.

$$\begin{aligned} \underset{p(y|x)}{\text{maximize}} \quad & H(p) = -\sum_x \tilde{p}(x) \sum_y p(y \mid x) \log p(y \mid x) \\ \text{subject to} \quad & E_{\tilde{p}}[f_i] - E_p[f_i] = 0 \quad 1 \leq i \leq F, \\ & \sum_y p(y \mid x) - 1 = 0 \quad \text{for all } x. \end{aligned} \tag{6}$$

We do not solve the above primal optimization problem directly. Instead, we derive an easier unconstrained dual optimization problem with the Lagrange method, where we find that $p(y \mid x)$ has a so-called "log-linear" parametric form:

$$p_{\boldsymbol{\lambda}}(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i \lambda_i f_i(x, y)\right) \quad \lambda_i \in \mathbb{R}$$

$$Z(x) = \sum_y \exp\left(\sum_i \lambda_i f_i(x, y)\right) \tag{7}$$

Parameter λ_i is the Lagrange multiplier corresponding to the equality constraint for f_i.

The dual objective function becomes:

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_x \tilde{p}(x) \sum_y \tilde{p}(y \mid x) \sum_i \lambda_i f_i(x, y) - \sum_x \tilde{p}(x) \log \sum_y \exp\left(\sum_i \lambda_i f_i(x, y) \right). \tag{8}$$

Therefore, ME estimation becomes the maximization for this $\mathcal{L}(\boldsymbol{\lambda})$. It can be shown that this is equivalent to the maximization of log-likelihood, which we define here as:

$$LL(\boldsymbol{\lambda}) = \log \prod_{x,y} p_{\boldsymbol{\lambda}}(y|x)^{\tilde{p}(x)\tilde{p}(y|x)}$$

The maximization of log-likelihood can easily be solved by using reasonable optimization algorithms since the log-likelihood is a concave function. For ME estimation, some specialized iterative algorithms such as the GIS algorithm (Darroch & Ratcliff, 1972) and the IIS algorithm (Della Pietra, Della Pietra, & Lafferty, 1997) have been applied. In addition, general-purpose gradient-based algorithms can be applied. Recently, Malouf (2002) compared several algorithms to estimate ME models including GIS, IIS, and gradient-based methods, and showed that the limited-memory variable metric (LMVM) (also known as L-BFGS), a quasi-Newton method that requires only limited memory size, converges much faster than other methods for NLP datasets.[3]

Our preliminary experiments also revealed that LMVM converges faster for our text categorization datasets. Because of this faster convergence, we could also achieve improved accuracy compared with GIS or IIS since we can achieve a solution that is much closer to the true optimum with the same computational cost. Therefore, we used LMVM to estimate the ME models that were to be compared with the inequality models. As we will describe later, we employed a variant of LMVM called BLMVM (Benson & Moré, 2001) to estimate the inequality ME models. This is because it supports the bounded parameters required in estimating inequality models, which cannot be handled by standard GIS or IIS.

LMVM (and BLMVM) only requires the function value and the gradient at a given point to perform optimization. Thus, we now only need to state the gradient of the objective function to specify the estimation.

The gradient of the objective function in Eq. (8) is computed as:

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

We can see that the original constraint $E_p[f_i] - E_p[f_i] = 0$ will certainly be satisfied at the optimal point.

2.2. *Gaussian MAP estimation*

In Gaussian MAP estimation (Chen & Rosenfeld, 1999, 2000; Johnson et al., 1999), we maximize the posterior probability of parameters given the training data instead of the likelihood of the training data, assuming that the prior distribution for the parameter of the log-linear model (Eq. (7)) is Gaussian centered around zero with variance σ_i^2. In this case, posterior probability can be represented as:

$$\begin{aligned} p_{\text{posterior}}(\boldsymbol{\lambda} \mid D) &\approx p(D|\boldsymbol{\lambda}) \times p_{\text{prior}}(\boldsymbol{\lambda}) \\ &\approx \prod_{x,y} p_{\boldsymbol{\lambda}}(y \mid x)^{\tilde{p}(x,y)} \times \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{\lambda_i^2}{2\sigma_i^2}\right). \end{aligned}$$

Taking the logarithm of the above and ignoring the constant terms, the objective function becomes the regularized log-likelihood:

$$\mathcal{L}(\boldsymbol{\lambda}) = LL(\boldsymbol{\lambda}) - \sum_i \left(\frac{1}{2\sigma_i^2}\right)\lambda_i^2. \tag{10}$$

The gradient becomes:

$$\frac{\partial \mathcal{L}(\boldsymbol{\lambda})}{\partial \lambda_i} = E_{\tilde{p}}[f_i] - E_p[f_i] - \frac{\lambda_i}{\sigma_i^2}. \tag{11}$$

At the optimal point, $E_{\tilde{p}}[f_i] - E_p[f_i] - \frac{\lambda_i}{\sigma_i^2} = 0$. Therefore, Gaussian MAP estimation can also be considered as relaxing equality constraints.

3. Inequality ME model

We will now describe the ME method with box-type inequality constraints (Eq. (3)). The derivation of the parametric form and the dual objective function is similar to that of the standard ME method. After we define the primal optimization problem, we use the Lagrange method.

Maximum entropy estimation with box-type inequality constraints (Eq. (3)) can be formulated as the following optimization problem:

$$\underset{p(y|x)}{\text{maximize}} \quad H(p) = -\sum_x \tilde{p}(x) \sum_y p(y \mid x) \log p(y \mid x), \tag{12}$$

$$\text{subject to} \quad E_{\tilde{p}}[f_i] - E_p[f_i] - A_i \leq 0 \quad \text{(upper constraints), and} \tag{13}$$

$$E_p[f_i] - E_{\tilde{p}}[f_i] - B_i \leq 0 \quad \text{(lower constraints).} \tag{14}$$

The box-type inequality constraints are reasonable in the following sense. First, if we let A_i,$B_i \to 0$, the problem is as close as standard ME estimation. Second, if

standard ME estimation has a solution, inequality ME estimation will also have a solution, since the feasible region—the region where all the constraints are satisfied—is not empty because, if equality constraints are satisfied, the inequality constraints are also satisfied.

By using the Lagrange method for optimization problems with inequality constraints, the following parametric form is derived (see Appendix A for details on derivation).

$$p_{\boldsymbol{\alpha},\boldsymbol{\beta}}(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i (\alpha_i - \beta_i) f_i(x, y) \right), \alpha_i \geq 0, \quad \beta_i \geq 0, \tag{15}$$

where α_i and β_i are the Lagrange multipliers (α_i corresponds to the upper inequality constraint (Eq. (13)) and β_i corresponds to the lower inequality constraint (Eq. (14)). Note that although each parameter must not be negative, $\alpha_i - \beta_i$ can be any real value (as we will describe below, $\alpha_i = 0$ when $\beta_i > 0$ and $\beta_i = 0$ when $\alpha_i > 0$). Thus, we can interpret that α_i is responsible for the positive importance of the feature while β_i is responsible for the negative. Although there are two parameters per feature, we can view the above model as a standard ME model by considering it as $\lambda_i = \alpha_i - \beta_i$ once the parameters are estimated. However, estimation is affected by the fact that there are two bounded parameters.

The dual objective function becomes:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = & - \sum_x \tilde{p}(x) \log \sum_y \exp\left(\sum_i (\alpha_i - \beta_i) f_i(x, y) \right) \\ & + \sum_x \tilde{p}(x) \sum_y \tilde{p}(y \mid x) \sum_i (\alpha_i - \beta_i) f_i(x, y) \\ & - \sum_i \alpha_i A_i - \sum_i \beta_i B_i. \end{aligned} \tag{16}$$

Then, the estimation is formulated as:

$$\underset{\alpha_i \geq 0, \beta_i \geq 0}{\text{maximize}} \ \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}). \tag{17}$$

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

$$LL(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \log \prod_{x,y} p_{\boldsymbol{\alpha},\boldsymbol{\beta}}(y \mid x)^{\tilde{p}(x,y)}.$$

Instead, we maximize:

$$LL(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \sum_i \alpha_i A_i - \sum_i \beta_i B_i. \tag{18}$$

The significant difference between inequality ME estimation and Gaussian MAP estimation is that the parameters are stabilized linearly in the inequality ME model (by penalty term $-\sum_i \alpha_i A_i - \sum_i \beta_i B_i$ as in Eq. (18)), while they are stabilized quadratically in Gaussian MAP estimation (by penalty term $-\sum_i (1/2\sigma_i^2)\lambda_i^2$ as in Eq. (10)). In this sense, inequality ME estimation penalizes large weights less and penalizes small weights more than Gaussian MAP estimation.

Since the objective function is still concave, we can use gradient-based optimization methods to solve this dual problem if they support bounded parameters. In this study, we used the BLMVM algorithm (Benson & Moré, 2001), a variant of the limited-memory variable metric (LMVM) algorithm, which supports bound constraints.[4]

The gradient of the objective function in Eq. (16) is:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_i} = E_{\tilde{p}}[f_i] - E_p[f_i] - A_i$$
$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_i} = E_p[f_i] - E_{\tilde{p}}[f_i] - B_i. \tag{19}$$

3.1. *Solution sparseness*

The existence of a sparse solution is predicted from the conditions in the Lagrange method with inequality constraints. The Karush-Kuhn-Tucker (KKT) conditions state that, at the optimal point (i.e., after training),

$$\alpha_i(E_{\tilde{p}}[f_i] - E_p[f_i] - A_i) = 0 \quad \text{and} \tag{20}$$
$$\beta_i(E_p[f_i] - E_{\tilde{p}}[f_i] - B_i) = 0. \tag{21}$$

These conditions mean that the equality constraint is maximally violated (i.e., $E_{\tilde{p}}[f_i] - E_p[f_i] = A_i$ or $-B_i$) when the parameter is non-zero, but, on the other hand, the parameter must be zero when the violation is strictly less than the width (i.e., $-B_i < E_{\tilde{p}}[f_i] - E_p[f_i] < A_i$)). Then, a feature after training can be classified as one of the following cases.

1. $\alpha_i > 0$ and $\beta_i = 0$ (*upper active*),
2. $\alpha_i = 0$ and $\beta_i > 0$ (*lower active*), or
3. $\alpha_i = 0$ and $\beta_i = 0$ (*inactive*).

When $\alpha_i - \beta_i \neq 0$ (i.e., cases 1 and 2), we say that the feature is *active*.[5] Only active features have an impact on the model's behavior. Inactive features can be removed from the model without changing its behavior. The case for $\alpha_i > 0$ and $\beta_i > 0$ is excluded since it is theoretically impossible because of definitions $A_i > 0$ and $B_i > 0$.

The sparse solution results from the existence of case 3. If many features fall into case 3 as a result of optimization, we say that the solution is sparse. There is a basic relation between widths and solution sparseness: the solution becomes more sparse as the widths

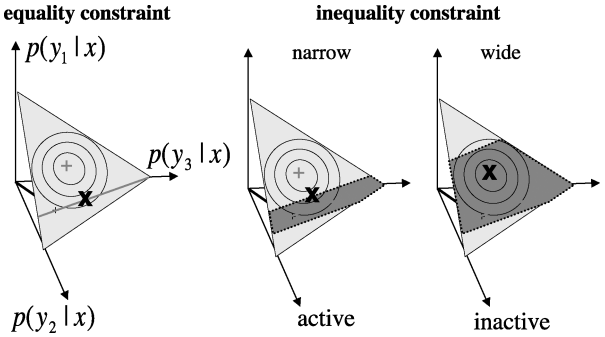


Figure. 1 The triangular plane represents $\sum_y p(y \mid x) - 1 = 0$ for a certain x. The entropy varies on this plane concavely with the maximum at the center marked '+'. The shaded region represents the feasible region for each case. The left-most case is the standard ME model (equality constraints); the middle is the inequality model with narrow widths; the right-most case is the inequality model with broad widths. The point marked 'x' indicates the estimated point.

increase. Figure 1 illustrates this in a very simplified way. As we can see from the figure, when the widths are sufficiently narrow, the feasible region does not contain the point at which entropy achieves the global maximum. Thus, the distribution selected by the optimization (indicated by X in the figure) will be on the edge of the feasible region (i.e., $E_{\tilde{p}}[f_i] - E_p[f_i] = A_i$ or $-B_i$). This means that the feature is active. When the widths are broad enough on the other hand, the feasible region might contain the global maximum entropy point. In this case, the chosen distribution is the global maximum entropy point itself. This means that the feature is inactive. As the widths A_i and B_i determine the width of the feasible region, it can be said that the broader the widths, the larger this feasible region is. The empirical relation between the widths and the sparseness of the solution will be demonstrated in the experiment.

Although there is this tendency, Figure 1 also suggests that solution sparseness is also affected by the distance between the global maximum entropy point and the feasible region. In addition, although it seems simple to predict the activity of a feature when we have only one feature as in Figure 1, the interaction between features complicates the matter. That is, there are cases where a feature that is active if it is alone becomes inactive when another feature is included in the model (and vice versa). There are also cases where a feature that is upper active alone becomes lower active with another feature (and vice versa). This feature interaction, however, is the reason for our devising inequality ME estimation. If the activity of features is determined independently beforehand, there is no need for running inequality ME estimation in terms of selecting features. However, the fact is that we have many interacting features in a model and the interactions between them should be solved by inequality ME estimation. Solution sparseness is a result of optimization (Eq. (17)), which considers all these factors.

However, note that a feature with zero empirical expectations (zero count) must be lower active if it is active, regardless of feature interaction. Because $E_p[f_i] \geq 0$ (since $f_i(x,y) \geq 0$) and $A_i > 0$, it is impossible to satisfy upper equality $E_{\tilde{p}}[f_i] - E_p[f_i] - A_i =$

0 when $E_{\tilde{p}}[f_i] = 0$. Thus, a feature with zero empirical expectation cannot be upper active. Lower equality, on the other hand, can be satisfied even if empirical expectation is zero. This is similar to Gaussian MAP estimation, where the weights for zero-count features must be zero or negative, as implied by the relation in Eq. (2).

4. Calculation of constraint width

This section describes how to determine constraint widths, A_i and B_i. Because our aim was to improve the generalization performance of the model, one straightforward way was cross-validation or development-set testing that estimate the best widths empirically. However, it is computationally impossible to do such tests for all combinations of possible widths. Therefore, we sought methods that determine the widths in a principled way with a few control parameters, for which finding the best values through development-set testing gives the widths close to the ideal ones. It is reasonable to think that the widths, A_i and B_i, are widened depending on the unreliability of the feature, because we can expect that unreliable features will become increasingly more inactive, and this will improve generalization performance. In this paper, we examine two methods that fulfill this expectation with a few (or one) control parameters.

4.1. Single width

The first is to use a common width for all features, which is calculated with the following formula.

$$A_i = B_i = W \times \frac{1}{L}, \tag{22}$$

where W is a constant *width factor* to control the widths, and L is the number of training examples. This method takes the reliability of training examples as a whole into account, and is called *single*. Note that although we can use different widths for A_i and B_i, we have restricted ourselves throughout the paper to using the same width for upper and lower inequality (i.e., $A_i = B_i$).

4.2. Bayesian width

The second, which we call *bayes*, is a method that determines widths based on a Bayesian framework that provides different widths for each feature based on their reliabilities.

The reasoning behind *bayes* is as follows. For many NLP applications including text categorization, we use features with the following form.

$$f_{i,j}(x, y) = h_i(x) \quad \text{if } y = y_j, \quad 0 \ \text{ otherwise.} \tag{23}$$

Even if two features have the same empirical expectation (count), the meaning differs depending on the frequency of the history $h_i(x)$. For example, $h_i(x)$ typically represents the

occurrence of word w_i in text categorization. If the frequency of the word is low, then the reliability of the corresponding feature will also be low. However, if the frequency of the word is high and the frequency of the feature is low, the feature is reliably having a negative importance.

To calculate widths that increase when the frequency of the history decreases, we use Bayesian estimation here. If we assume approximation, $\tilde{p}(y|x) \approx \tilde{p}(y|h_i(x) > 0)$, the empirical expectation can be interpreted as follows.[6]

$$E_{\tilde{p}}[f_{i,j}] = \sum_{x:\, h_i(x)>0} \tilde{p}(x)\tilde{p}(y = y_j \mid h_i(x) > 0)h_i(x). \tag{24}$$

Here, the source of unreliability is $\tilde{p}(y \mid h_i(x) > 0)$. We consider $\tilde{p}(y = y_j \mid h_i(x) > 0)$ to be parameter θ for Bernoulli trials. That is, $p(y = y_j \mid h_i(x) > 0) = \theta_{i,j}$ and $p(y \neq y_j \mid h_i(x) > 0) = 1-\theta_{i,j}$. We estimate the posterior distribution of θ from the training examples by Bayesian estimation and utilize variance to calculate the width. With uniform distribution as a 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.

$$V[\theta] = \frac{(1+k)(1+n-k)}{(2+n)^2(n+3)}. \tag{25}$$

Letting $k = c(f_{i,j}(x, y) > 0)$ and $n = c(\ h_i(x) > 0\ )$, we obtain fine-grained variances narrowed according to $c(h_i(x) > 0)$. Assuming the independence of training examples, the variance in empirical expectation can be calculated as:

$$V[E_{\tilde{p}}[f_{i,j}]] = \left[\sum_{x:\, h_i(x)>0} \{\tilde{p}(x)h_i(x)\}^2\right] V[\theta_{i,j}]. \tag{26}$$

We then calculate the widths as follows.

$$A_i = B_i = W \times \sqrt{V[E_{\tilde{p}}[f_{i,j}]]}. \tag{27}$$

As with the *single* method, W is the only control parameter.

5. Soft width extensions

This section presents extensions of the inequality ME model, which we call *soft-width* extensions. The soft-width extension 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 for SVMs, and in fact, the mathematical derivation is similar.[7] If we penalize the slack variables by the square of their L_2 norm, we obtain a natural combination of the inequality ME model and Gaussian MAP estimation. We refer to this extension, using the L_2 penalty, as the *2-norm inequality ME* model. As Gaussian MAP estimation has been shown to be successful in several NLP tasks, it should

be interesting empirically as well as theoretically to integrate Gaussian MAP estimation into inequality ME estimation. In addition to the 2-norm extension, we can see what happens when we penalize the slack variables by their L_1 norm (*1-norm inequality ME model*) as in 1-norm soft-margin extension for SVMs.

5.1. 2-norm penalty extension

Our 2-norm extension to the inequality ME model is formulated as follows.

$$\underset{p(y|x),\boldsymbol{\delta},\boldsymbol{\gamma}}{\text{maximize}} \quad H(p) - C_1 \sum_i {\delta_i}^2 - C_2 \sum_i \gamma_i^2, \tag{28}$$

$$\text{subject to} \quad E_{\tilde{p}}[f_i] - E_p[f_i] - A_i \le \delta_i \quad \text{and}$$
$$E_p[f_i] - E_{\tilde{p}}[f_i] - B_i \le \gamma_i, \tag{29}$$

where C_1 (>0) and C_2 (>0) are the penalty constants.[8,9]

This formulation can be viewed as an extension inspired by the 2-norm soft-margin extension of SVM. At the same time, it can be seen as an extension inspired by Gaussian MAP estimation. As Chen and Rosenfeld (2000) pointed out, Gaussian MAP estimation can be viewed as a fuzzy ME model, whose primal formulation is:

$$\underset{p(y|x),\boldsymbol{\delta}}{\text{maximize}} \quad H(p) - \sum_i \frac{\sigma_i^2}{2} \delta_i^2$$
$$\text{subject to} \quad E_{\tilde{p}}[f_i] - E_p[f_i] = \delta_i \tag{30}$$

Thus, we can notice a similarity between this 2-norm ME estimation and Gaussian MAP estimation in that the slack variables are penalized by the square of their L_2 norm.

The parametric form of 2-norm extension is identical to the inequality ME model (Eq. (15)). However, the dual objective function becomes:

$$LL(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \sum_i \left(\alpha_i A_i + \frac{\alpha_i^2}{4C_1} \right) - \sum_i \left(\beta_i B_i + \frac{\beta_i^2}{4C_2} \right). \tag{31}$$

Accordingly, the gradient becomes:

$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_i} = E_{\tilde{p}}[f_i] - E_p[f_i] - \left(A_i + \frac{\alpha_i}{2C_1} \right),$$
$$\frac{\partial \mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_i} = E_p[f_i] - E_{\tilde{p}}[f_i] - \left(B_i + \frac{\beta_i}{2C_2} \right). \tag{32}$$

We can see that this model is a natural combination of the inequality ME model and Gaussian MAP estimation. It is important to note that the possibility of a sparse solution is preserved in the 2-norm extension above because of inequalities in the constraints.

5.2. *1-norm penalty extension*

It is also possible to impose 1-norm penalties in the objective function, as in the 1-norm soft-margin extension of SVM (Cristianini & Shawe-Taylor, 2000). If we impose 1-norm penalties, we obtain the following optimization problem.

$$\begin{aligned} \underset{p(y|x),\boldsymbol{\delta},\boldsymbol{\gamma}}{\text{maximize}} \quad & H(p) - C_1 \sum_i \delta_i - C_2 \sum_i \gamma_i, \\ \text{subject to} \quad & E_{\tilde{p}}[f_i] - E_p[f_i] - A_i \leq \delta_i \ \ (\delta_i > 0), \quad \text{and} \\ & E_p[f_i] - E_{\tilde{p}}[f_i] - B_i \leq \gamma_i \ \ (\gamma_i > 0). \end{aligned}$$

The parametric form and dual objective function for this optimization problem are identical to those of the inequality ME model, except that the parameters are also upper-bounded as $0 \leq \alpha_i \leq C_1$ and $0 \leq \beta_i \leq C_2$. That is, excessive parameters are explicitly prevented by the bounds. We will not evaluate this 1-norm extension in the experiments and leave this for future research.

6. Related work

Our work is not the first that has applied inequality constraints for ME estimation. As we previously mentioned, Khudanpur (1995) suggested inequality constraints $a_i \leq E_p[f_i] \leq b_i$ and showed that standard iterative scaling such as GIS and IIS cannot be used to train the ME models derived from these inequality constraints. Newman (1977) presented an ME model with inequality constraints with the form $\sum_i W_i (E_{\tilde{p}}[f_i] - E_p[f_i])^2 \leq \sigma^2$, for spectral analysis. However, this model is for continuous distributions, and there is only one constraint (i.e., constraints are not for each feature) and weights W_i instead seem to work as the λ_i of standard ME models. It is unclear how to apply this model to NLP tasks, where discrete distributions are used. Fang et al. (1997) contains more general definitions and references to ME estimation using inequality constraints. However, as Chen and Rosenfeld (2000) noted, ME estimation using inequality constraints has not been applied or evaluated for NLP.

We recently noted that Goodman (2003, 2004) derived an ME model, which is very closely related to our inequality ME model. Goodman used an exponential prior, which is written as follows, instead of a Gaussian prior.

$$p(\lambda_i) = a_i \exp(-a_i \lambda_i) \ \ \text{(nonzero only for } \lambda_i > 0)$$

He proposed the use of an exponential prior from an observation of the actual distributions of parameters, which indicated that the Gaussian is not necessarily the best prior. With an exponential prior, the objective function becomes: $\mathcal{L}(\boldsymbol{\lambda}) = LL(\boldsymbol{\lambda}) - \sum_i a_i \lambda_i$. The model has the bounds of parameters: $\lambda_i > 0$. He also noted that an exponential prior favors zero parameters, which is not the case with the Gaussian prior. Therefore, it is clear that Goodman's exponential prior model is almost the same as our inequality ME model. However, there are slight differences and these are as follows. First, his model has only upper

active features. That is, Goodman's model is equivalent to our inequality ME model where the lower constraints (Eq. (14)) are not imposed at all. This one-sided model might have an advantage in terms of training speed since it does not double the number of parameters, achieving the desired improvement in accuracy. However, as Goodman himself pointed out, this one-sided model discards some information when negative weights are appropriate.[10] In the discussion on estimation cost, we will briefly compare the benefit of one-sided and double-sided inequality ME models. Second, the bounds $\lambda_i > 0$ are rather artificial in Goodman's model, while they were given a natural interpretation as the consequence of inequality constraints in our model. In addition, his training algorithm is a variant of GIS modified to support the parameter bounds ($\lambda_i > 0$). However, GIS is no longer state-of-the-art in estimating ME models. Although he pointed out difficulties with the gradient-based method, our experiments revealed that BLMVM estimates inequality models very well. Last, he did not focus much on solution sparseness and did not demonstrate how sparse his model actually becomes. Although solution sparseness is not the direct purpose of study, and depends on the task, how sparse the solution becomes is of considerable interest since it indicates the essential number of features for that task. Our experiments show that for the text categorization task, many features become inactive when generalization performance is maximized. Despite these differences, his study reveals the relation between our inequality ME models and MAP estimation.

7. Experiments

This section describes a series of experiments that demonstrate the properties and advantages of inequality ME models. The experiments were conducted using the text categorization task.

7.1. Experimental setting

We used the "Reuters-21578, Distribution 1.0" dataset and the "OHSUMED" dataset for the experiments.

The Reuters dataset developed by David D. Lewis is a collection of labeled newswire articles.[11] We adopted "ModApte" split to divide the collection into training and test sets. We obtained 7,048 documents for training, and 2,991 documents for testing. We further divided the test set into another two sets. The first half (1,496 documents) was used as the development set to tune the control parameters, and the second half (1,495 documents) was left for the final evaluation using tuned control parameters. 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 "$A/B/C \mapsto A$", so that the major part of the MeSH term would be used as a target category. We considered 100 such simplified terms that occurred most frequently in the training set as the target categories. We extracted 9,947 abstracts for training, and 9,948 abstracts for testing from the file "ohsumed.91" in the OHSUMED collection. We

further divided the test set into a development set (4,974 documents) and an evaluation set (4,974 documents), as we did for the Reuters collection.

After the stop words had been removed and all the words had been downcased (no stemming was performed), documents were converted to bag-of-words vectors with TFIDF values (Salton & Buckley, 1988), which are well-used representations for text categorization. TFIDF is a product of *term frequency* ($TF(w_i)$), which is the number of times word w_i occurs in the document and *inverse document frequency* ($IDF(w_i)$), which is calculated as

$$IDF(w_i) = \log \frac{|D|}{\text{number of documents where } w_i \text{ occurs}},$$

where D denotes all the documents considered. The idea behind IDF is that a word that rarely occurs is potentially important, and the usefulness of TFIDF has been proved in many IR-related tasks, including text categorization. The IDF value was calculated using the training set.[12] Then the document vector was normalized by the L_1 norm so that the sum of all elements became 1.

Since the text categorization task requires that multiple categories be assigned if appropriate, we constructed a binary categorizer $p_c(y \mid d)$ for each category c, where $y \in \{+1,-1\}$ ($+1$ means that category c should be assigned for document d, and -1 that it should not). If $p_c(+1|d)$ is greater than 0.5, category c is assigned. To construct a conditional maximum entropy model, we used the feature functions with the form in Eq. (23), where $h_i(d)$ returns the TFIDF value of the i-th word of the document vector. Note that, when we considered both $f_{+1,i}(y,d)$ and $f_{-1,i}(y,d)$ for all $h_i(d)$ that occurred at least once in the training set, the average number of features in a categorizer (averaged over all categories) became 63,150.0 for the Reuters dataset, and 116,452.0 for the OHSUMED dataset. These numbers are the same when the cut-off threshold is zero with the cut-off method, and we will compare these with the number of active features in the inequality models later.

We implemented the estimation algorithms by extending an ME estimation tool, Amis.[13] As we discussed, we used LMVM to estimate the standard ME models. We also used it to estimate the Gaussian MAP models. The only difference is in the gradient we gave to the LMVM module. The gradient in Eq. (9) was used for standard ME models, and the gradient in Eq. (19) was used for Gaussian MAP models. For inequality ME estimation, in addition to setting the parameter bounds $\alpha_i, \beta_i \geq 0$ for the BLMVM module, we added a hook that checks the KKT conditions after the normal convergence test. Specifically, we examine whether the following conditions hold for all features.

$$\frac{E_{\tilde{p}}[f_i] - E_p[f_i] - A_i}{A_i} \leq T \quad \text{if } \alpha_i = 0, \quad \frac{|E_{\tilde{p}}[f_i] - E_p[f_i] - A_i|}{A_i} \leq T \quad \text{if } \alpha_i > 0,$$
$$\frac{E_p[f_i] - E_{\tilde{p}}[f_i] - B_i}{B_i} \leq T \quad \text{if } \beta_i = 0, \quad \frac{|E_p[f_i] - E_{\tilde{p}}[f_i] - B_i|}{B_i} \leq T \quad \text{if } \beta_i > 0,$$

where T is the tolerance for the KKT checks.

LMVM and BLMVM are provided in the Toolkit for Advanced Optimization (TAO) (Benson et al., 2002), which is well established and was also used by Malouf (2002). By using algorithms in the same family implemented in the same code base, we could compare the inequality ME model with the standard and Gaussian ME models on an even basis. Comparison also has a practical value because each model's accuracy will be the highest possible with the existing optimization algorithms for ME estimation.

The tolerance for the normal convergence test (relative improvement in the objective function value) and the KKT check was 10^{-4}. Although we noted that a feature cannot be upper and lower active at the same time theoretically, this is not exactly achieved because we performed numerical optimization. In actual implementation, we set a small tolerance for the existence of features that were upper and lower active at the same time. That is, we stopped the training if the KKT check failed many times and the ratio of 'bad' (upper and lower active at the same time) features in active features was lower than 0.01.

7.2. Comparison with standard and Gaussian MAP ME estimation

We first compared inequality ME models (with no cut-off) with standard ME models and Gaussian MAP ME models (both with cut-off) to demonstrate the properties and advantages of inequality ME models. We compared the following models:

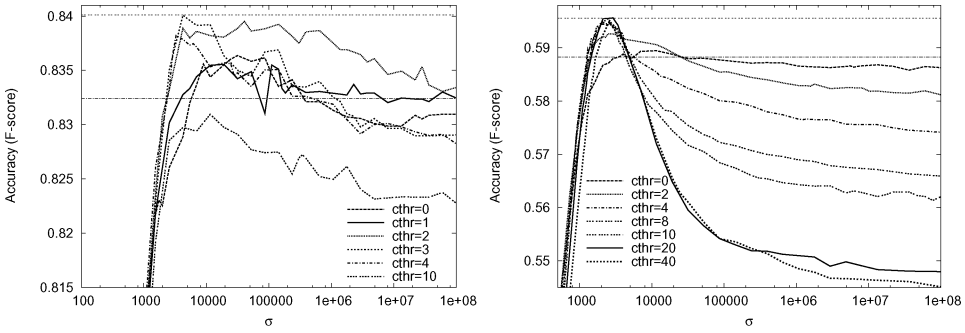
- ME models only with cut-off (*cut-off*)
- ME models with cut-off and Gaussian MAP estimation (*gaussian*)
- Inequality ME models (*ineq*)
- Inequality ME models with 2-norm extension (*2-norm*)[14]

We compared the two methods to determine the widths, *single* and *bayes*, for the inequality ME models, as described in Section 4.

To test practical generalization performance, we first found the best values for the control parameters, using the development set, and then observed how these values worked in the evaluation set. That is, how good control parameters can be discovered is part of the performance of the model. Throughout the experiments, performance was assessed by micro-averaged recall, precision, and F-score, which are widely used for the text categorization task. Letting Cor_c be the number of documents where the system assigned category c correctly, Std_c be the number of documents where c should be assigned according to the correct labeling, and Sys_c be the number of documents where the system assigned c, these measures could be defined as follows.

$$\text{Recall} = \frac{\sum_c Cor_c}{\sum_c Std_c}, \quad \text{Precision} = \frac{\sum_c Cor_c}{\sum_c Sys_c}, \quad F = \frac{2\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

We considered the value with which the micro-averaged F-score marked the highest value for the development set as the best value for the control parameter. We searched for the best control parameter by exhaustively testing several values within a given range.[15] Note that we considered the values of the control parameters were the same over all categories



a. Reuters b. OHSUMED

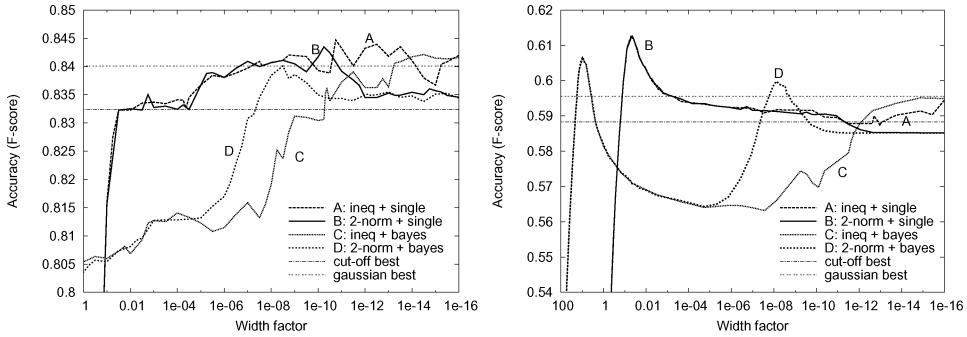
Figure. 2 Best *cthr* and σ combinations for Gaussian MAP estimation.

to simplify the experiments. Therefore, the accuracies can be higher than those reported in our experiments if the control parameters are optimized for each category independently.

The control parameter for *cut-off* is the count threshold *cthr*. Features that occurred $\geq$ *cthr* times in the training set were included in the model. We tested *cthr*: 0, 1, 2, . . ., 10, 12, 15, 20, 30, 40, 50, 60, . . ., and found that $cthr = 2$ was the best for the Reuters dataset, and $cthr = 0$ was the best for the OHSUMED dataset.

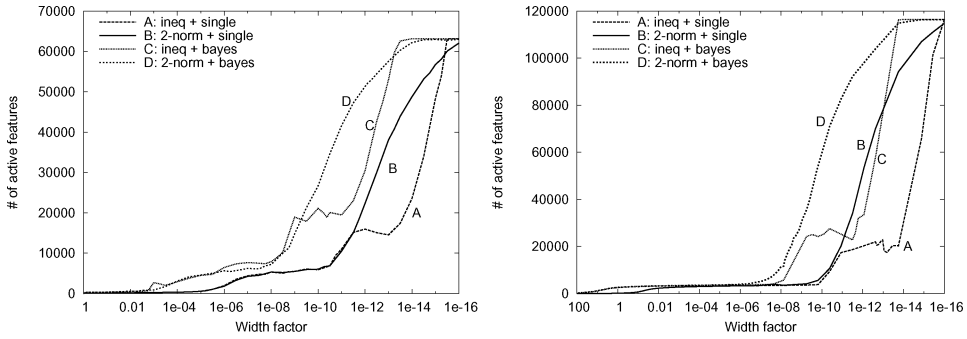
The control parameters for the *gaussian* were *cthr* and σ of a Gaussian prior. Although we can use a different σ_i for each feature in Gaussian MAP estimation, we used common variance σ for the *gaussian*. Thus, the *gaussian* roughly corresponds to *single* in the way it deals with unreliability of features. We varied σ within the range for each *cthr*. Figure 2 plots the results of searching for the best *cthr* and σ. We found that $cthr = 3$ and $\sigma = 4.22 \times 10^3$ were the best for the Reuters dataset, and that $cthr = 20$ and $\sigma = 2.90 \times 10^3$ were the best for the OHSUMED dataset.[16] The lower horizontal line indicates the highest accuracy for *cut-off*. As the previous work indicated (Nigam, John, & McCallum, 1999), Gaussian MAP estimation outperformed standard ME estimation in text categorization. The curious point here is that count thresholds greater than the best thresholds for standard ME estimation yielded the best accuracies, since it seems plausible to expect that MAP estimation would allow the count threshold to decrease because of its regularization effect. However, this is probably more complicated. Although we are still not sure why this behavior was observed, one possible explanation is that we should also count the (counter intuitive) properties that estimation tends to overfit more with less features because of faster convergence (if the convergence levels are the same). Gaussian MAP estimation enables more stable estimation with such less features compared with standard ME estimation due to its regularization.

The control parameter for inequality ME models was width factor W. Figure 3 plots accuracies of inequality ME models for various W. In this experiment, we started our estimates with all possible features (i.e., $cthr = 0$) and relied on the ability of inequality models to remove unnecessary features through solution sparseness. The horizontal lines indicate the best accuracies for *cut-off* and *gaussian* previously discussed. We can see that the inequality models outperform standard ME estimation and Gaussian MAP estimation with an appropriate value for W in both datasets. As can be seen, the OHSUMED dataset seems



a. Reuters b. OHSUMED

Figure. 3 Relation between width factor W and accuracy of inequality ME models for development set.



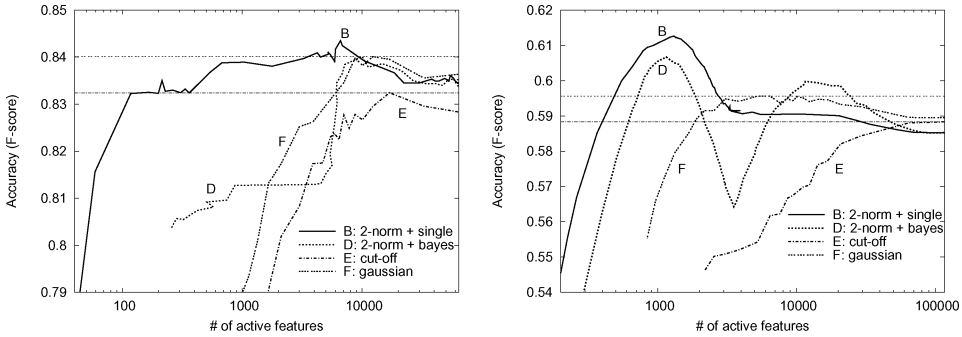
a. Reuters b. OHSUMED

Figure. 4 Relation between width factor and average number of active features.

more difficult than the Reuters. However, the improvement in the OHSUMED is greater than that in the Reuters dataset. The plotted curves are smoother for the OHSUMED dataset than they are for the Reuters. This may be because the development set for the Reuters dataset is relatively small (1,496 documents) compared with that of the OHSUMED (4,974 documents). We could not observe any advantage in using the *bayes* method to determine widths. Although 2-norm extension seemed to increase the accuracies for *bayes*, we could not observe apparent advantages in terms of greatest accuracy.

Figure 4 plots the average number of active features (averaged over all categories) for each inequality ME model for various width factors. The active features increased when the widths narrowed, as expected.

Figure 5 plots the accuracy of each model for the development sets as a function of the number of active features. For *cut-off* and *gaussian*, the best accuracy for each count threshold is plotted against the number of features with that count threshold. Only *2-norm* is displayed for inequality models to make the curves easier to distinguish. We can see that inequality ME models achieve the highest accuracy with a fairly small number of active



a. Reuters b. OHSUMED

Figure. 5 Accuracies of inequality models plotted as function of average number of active features.

Table 1. Comparison of accuracies for Reuters dataset.

Methods	Best control parameters	Active features	Accuracy (dev) P/R/F	Accuracy (eval) P/R/F
cut-off	c = 2	16,961.9	87.06/79.75/83.24	90.11/82.95/86.38
gaussian	c = 3, $\sigma = 4.22e3$	12,326.6	91.21/77.86/84.01	94.18/80.90/87.04
ineq+single	$W = 1.78e{-}11$	9,479.9	87.87/81.32/84.47	90.63/84.40/87.41
2-norm+single	$W = 5.62e{-}11$	6,611.1	87.98/81.01/84.35	91.22/84.23/87.59
ineq+bayes	$W = 3.16e{-}15$	63,150.0	89.54/79.49/84.21	92.98/82.40/87.37
2-norm+bayes	$W = 3.16e{-}9$	10,022.3	89.02/79.54/84.01	92.47/83.18/87.57

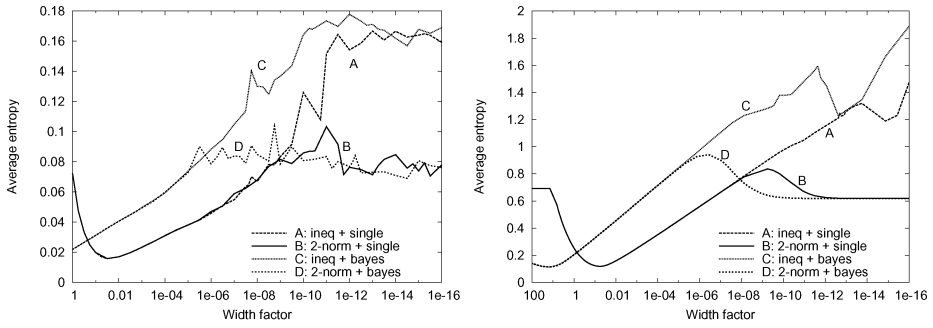
features, by removing unnecessary features through solution sparseness. Also, inequality models (*single* width) consistently achieve much higher accuracies than *cut-off* and *gaussian* with a small number of features, while *cut-off* and *gaussian* considerably degrade accuracy as count threshold increases.

Tables 1 and 2 summarize the results thus far and show how well each model performs for an evaluation set with the found best control parameters. Inequality ME models outperform other models for both the development and evaluation sets. That is, the inequality ME model is generally superior to standard ME estimation and Gaussian MAP estimation. The features reduced by inequality ME models are considerable. For the OHSUMED dataset, the features were reduced by a factor of 100 compared with all possible features, and by over a factor of 4 compared with the best Gaussian MAP estimates.

We can see that 2-norm extension has an advantage in being robust. That is, 2-norm models outperformed normal inequality models in the evaluation set. Of course, the level of difference here cannot be proved to be statistically significant only from these experiments. However, to see whether there was actually a difference between normal inequality ME estimation and the 2-norm extension, we investigated the relation between width factor W and the averaged cross entropy of each inequality model for the development set. Here, we used cross entropy as another measure for the behavior of probabilistic models. The average cross entropy was calculated as $-\frac{1}{C}\sum_c \frac{1}{L}\sum_i \log p_c(y_i \mid d_i)$, where C is the number of

Table 2. Comparison of accuracies for OHSUMED dataset.

Methods	Best control parameters	Active features	Accuracy (dev) P/R/F	Accuracy (eval) P/R/F
cut-off	$c = 0$	116,452.0	68.37/51.63/58.83	68.12/51.04/58.35
gaussian	$c = 20, \sigma = 2.90e3$	5,252.3	69.54/52.09/59.56	69.92/51.35/59.21
ineq+single	$W = 4.81e{-}2$	1,257.6	72.49/53.08/61.28	72.82/52.73/61.17
2-norm+single	$W = 4.50e{-}2$	1,316.5	72.15/53.23/61.26	72.54/52.96/61.23
ineq+bayes	$W = 9.46$	1,136.6	72.66/52.05/60.65	72.56/51.60/60.31
2-norm+bayes	$W = 9.46$	1,154.5	72.65/52.08/60.67	72.60/51.89/60.32



a. Reuters b. OHSUMED

Figure. 6 W vs. average cross entropy for development set.

categories. Figure 6 plots the results. The cross entropy of 2-norm models is clearly different and more stable than that of normal inequality models. This difference is one explanation for the robustness of 2-norm extension, although there is no absolute relation between accuracy and cross entropy (e.g., the best accuracy for the Reuters dataset is not achieved with the lowest entropy). This stability in 2-norm extension was also observed in Figure 4, which plotted the relation between W and the number of active features as smoother curves. This smoothness might avoid the development set from being overfitted through tuning W.

7.3. Performance with more sparse features

In this experiment, we investigated using the OHSUMED dataset how inequality ME models perform when more sparse features are included. We here consider the case where we include bi-gram features, which use two consecutive words as an unit.

Although intuitively reasonable, the empirical usefulness of beyond-one-word features has been controversial (Lewis, 1992; Mladenic & Globelnik, 1998; Scott & Matwin, 1999; Tan, Wang, & Lee, 2002; Pang, Lee, & Vaithyanathan, 2002), and n-gram features are still not standard for text categorization. One reason might be that sparseness caused by these

Table 3. Effect of adding bi-gram features (best results are listed).

Methods	Best control parameters	Active features	Accuracy (dev) P/R/F	Accuracy (eval) P/R/F
cut-off (uni)	c = 0	116,452.0	68.37/51.63/58.83	68.12/51.04/58.35
cut-off (+bi)	c = 4	74,183.0	69.11/53.18/60.11	69.43/52.52/59.81
gaussian (uni)	c = 20, $\sigma = 2.90e3$	5,252.3	69.54/52.09/59.56	69.92/51.35/59.21
gaussian (+bi)	c = 2, $\sigma = 4.47e6$	176,745.3	71.85/51.60/60.07	72.42/50.92/59.80
ineq (uni)	$W = 4.81e{-}2$	1,257.6	72.49/53.08/61.28	72.82/52.73/61.17
ineq (+bi)	$W = 1.15e{-}2$	1,547.1	72.17/53.62/61.53	72.87/53.36/61.61

features sometimes cancels out their benefits. Note that the maximum average number of features with bi-gram features in the OHSUMED dataset is 1,241,546.0, which is over 10 times greater than that with only uni-gram features (116,452.0). Another reason is that they increase the cost of both training and categorization. We can expect that inequality ME models prevent overfitting caused by including bi-gram features. In addition, inequality ME models would solve the problem of runtime costs because of sparse solutions in some cases.

In this experiment, we show that the inequality ME model achieves the highest accuracy of ME models even if bi-gram features are included. We use TFIDF for bi-gram features as well. TF for bi-gram $w_i\, w_j$ is the frequency of the two consecutive words in the document, and IDF is calculated as:

$$IDF(w_i w_j) = \log \frac{|D|}{\text{number of documents where } w_i w_j \text{ occurs}}.$$

Table 3 lists the results, comparing them with those of the models using only uni-gram features. We can see that the inequality ME model had the highest accuracy of ME models even if bi-gram features are included, although all ME models improved accuracy by including bi-gram features. Striking point here is the stable ability for feature selection of inequality ME estimation. The inequality ME estimation achieved improved accuracy just by adding a few features. Note that 511.7 features that were not active in the uni-gram model were newly added to the bi-gram model on average. Out of these 511.7 features, 372.1 features were bi-gram features (222.1 uni-gram features that had been active in the uni-gram model were inactive in the bi-gram model). On the other hand, standard ME estimation found fewer features are the best and Gaussian MAP estimation found much more features are the best. This indicates the unstable behavior of the combination of these estimation methods and cut-off thresholding.

7.4. Effect of count threshold for inequality ME model

In the previous experiment, we fixed the count threshold for inequality ME models at zero to check what effect solution sparseness had. In this experiment, we assessed what effect the count threshold had on inequality ME models. Figure 7 plots the W-accuracy curves

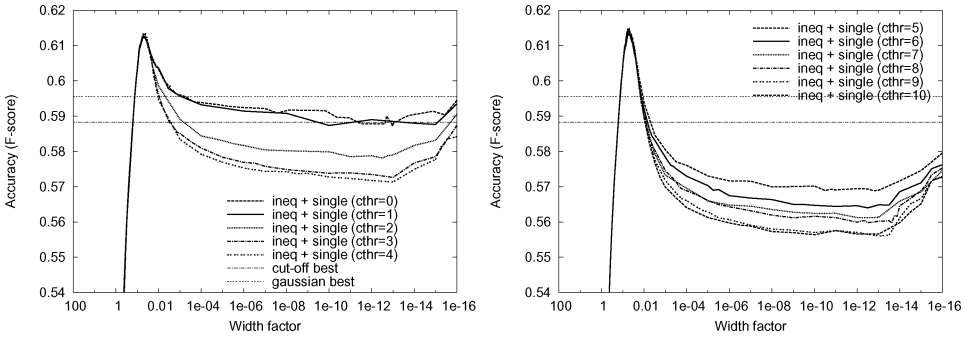


Figure. 7 Effect of count threshold for inequality ME model (OHSUMED dataset).

with various count thresholds for an *ineq+single* model in the OHSUMED dataset and Table 4 lists the details on the best accuracies. Although the count threshold seems to affect accuracy at moderate W, the best accuracy was not significantly affected (though thresholds such as 4 seem to yield the highest accuracy). This means that features with low counts are not that important in achieving high accuracies for the OHSUMED dataset. Inequality models perform such thresholding automatically. With cut-off thresholds around 15, however, the best accuracy started to decrease. This means that there are essential features for the OHSUMED dataset at around 15 counts. However, there are features that should be omitted to improve accuracy even in very frequent features. The inequality ME model omits such features appropriately. For example, the maximum number of features is 3856.8 with a count threshold of 30, but the number of active features after training is only 517.8. Note that Gaussian MAP estimation only achieved an F-score of 59.40 for the development set and 59.13 for the evaluation set at a count threshold of 30.

7.5. The cost of estimation

As we previously mentioned, the cost of estimating inequality models will be higher than that of standard ME models probably because of the inherent difficulty of bounded optimization and the doubled number of parameters. However, as can be seen from the number of experiments we successfully conducted, the additional estimation cost is probably tolerable for the sizes of the usual annotated NLP resources used for supervised training. Table 5 lists the times (I/O times excluded) required to estimate the best models (training all categories) with $cthr = 0$ in the first experiment on 1.27 GHz Pentium III machines. Gaussian MAP estimation requires much less time, reflecting the fact that it favors small weights and the estimates will reach the optimal point quicker. Inequality ME models certainly require much more time as expected.

One-sided inequality ME models (or equivalently exponential MAP ME models (Goodman, 2003, 2004)), which are derived using only the upper constraints, do not double the number of parameters. Therefore, it may be more efficient and it is desirable if it does not degrade the accuracy. To investigate the risk of not having lower active features and also to obtain greater insight into the causes of increased estimation costs, we briefly compared

Table 4. Best accuracies of inequality ME models (*ineq+single*) with various count thresholds (OHSUMED).

cthr	Maximum features	Active features	Accuracy (dev) P/R/F	Accuracy (eval) P/R/F
0	116,452.0	1,257.6	72.49/53.08/61.28	72.82/ 52.73/61.17
1	64,385.1	1,098.4	72.67/52.90/61.23	73.06/52.62/61.18
2	28,865.4	1,083.5	72.02/53.39/61.32	72.46/53.17/61.33
3	20,764.1	867.2	73.30/52.65/61.29	73.61/52.40/61.22
4	16,692.2	963.0	71.25/53.83/61.30	71.61/53.46/61.32
5	14,216.5	855.5	72.57/53.10/61.33	73.04/52.79/61.29
6	12,428.5	823.8	72.68/53.12/61.38	73.03/52.73/61.24
7	11,137.0	750.5	73.21/52.96/61.46	73.58/52.43/61.23
8	10,154.7	773.6	72.68/53.26/61.47	73.21/52.63/61.23
9	9,324.4	709.4	73.23/52.91/61.43	73.65/52.28/61.15
10	8,666.5	733.8	72.78/53.19/61.46	73.29/52.58/61.23
12	7,614.8	701.2	72.83/53.08/61.41	73.33/52.39/61.12
15	6,463.5	663.2	72.88/53.01/61.38	73.43/52.42/61.17
20	5,251.3	576.7	73.65/52.37/61.21	73.91/51.92/60.99
30	3,856.8	517.8	73.77/51.97/60.98	73.96/51.56/60.76

Table 5. Cost of estimating inequality ME models (time in seconds).

Dataset	*cut-off*	*gaussian*	*ineq+single*	*2-norm+single*
Reuters	1,624	518	8,151	7,409
OHSUMED	2,976	942	6,867	6,859

the one-sided inequality ME models with the double-sided inequality ME models we had used so far. For the Reuters dataset, the one-sided model achieved an 84.14 F-score for the development set and 87.10 for the evaluation set, with 8134.7 (average) active features, taking 5,720 seconds to estimate all categories. For the OHSUMED dataset, it achieved 61.22 for the development set and 61.13 for the evaluation set, with 541.6 active features, taking 7,658 seconds to estimate. These results prove that double-sided models (though the slight difference in the OHSUMED dataset) outperform one-sided models in terms of accuracy, indicating that lower active features contain some useful information. The one-sided models had fewer active features than the double-sided models as expected. In terms of training costs, we could not observe expected results. The training time was decreased for the Reuters dataset, as expected, but was increased for the OHSUMED dataset, as opposed to expectations, and in any case, the training for one-sided inequality ME models was far more expensive than that for standard ME and Gaussian MAP ME models. This suggests that the dominant cause of increased estimation costs for inequality ME estimation was not the doubled number of parameters, but more plausibly the inherent difficulty of bounded optimization.

Therefore, we could reduce the estimation costs, for example, by removing features when they are inactive for several iterations; this means removing bounded constraints from the estimation. In addition, if the doubled parameters are not the dominant cause, a great deal of additional time might be spent because of the KKT checks, e.g., more iterations due to the tightened convergence criterion and the actual time spent on the KKT check itself. We simply added KKT checks to our estimation algorithm, following the convention of optimization derived using the Lagrange method (e.g., SVM training), since our primal goal was not fast estimation but the evaluation of accuracies. Designing estimation so that KKT checks can be omitted will also improve estimation efficiency. In fact, the training time was decreased to 2,482 seconds for the Reuters dataset with increased F-scores of 84.53 (dev) and 87.78 (eval), and to 3,073 seconds for the OHSUMED with decreased F-scores of 61.21 (dev) and 61.00 (eval), if we did not check the KKT conditions at all. The results indicate the possibility of faster estimation, although the results also indicate an unpredictable effect of convergence criteria on the accuracy (we think this is a common problem in evaluating machine learning algorithms). Note that, however, the change in the accuracy does not override the advantages of inequality ME estimation.

Of course, we can omit many features through count thresholding before training without affecting the best accuracy in some cases, as the experiment in Section 7.4 demonstrated. This will also reduce the cost of estimation.

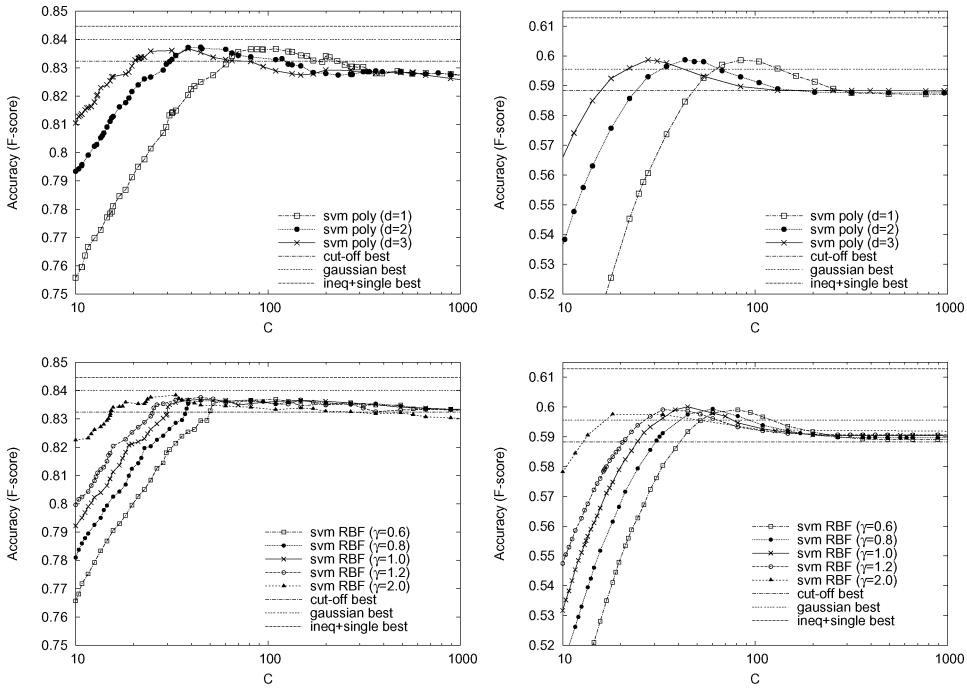
7.6. Comparison with Support Vector Machines

In our final evaluation, we compared inequality ME models with SVM-based text categorization. Support Vector Machines (SVMs) (Vapnik, 1995) are now considered a state-of-the-art machine learning method for many NLP tasks including text categorization. Since Joachims (1998b), there have been a number of previous studies that applied SVMs to text categorization and these demonstrated a high level of performance. Therefore, a comparison with an SVM-based categorizer would be considered a serious evaluation.

Data collections were the same as in the previous experiments. We used the polynomial kernel: $(sx{\cdot}y+r)^d$ with $s=1, r=1$ and the RBF kernel: $\exp(-\gamma \| x-y \|^2)$ as in Joachims (1998b). We evaluated text categorizers using these kernels for several ds of the polynomial kernel and several γs of the RBF kernel. Although there were many other candidates[17], soft-margin constant C (Cristianini & Shawe-Taylor, 2000) is most equivalent to the W and σ of ME models for the purposes of comparison. Therefore, we considered C as the control parameter to be tuned using the development set. We used SVM$^{\text{light}}$ (Ver. 5.0) (Joachims, 1998a) to estimate SVMs.

Figure 8 plots the results of searching for the best C for each kernel and dataset pair. Tables 6 and 7 list the best accuracies for SVM-based categorizers for each dataset. We can see that SVMs outperform at least standard ME models for both datasets, and outperform Gaussian MAP models for the OHSUMED dataset, proving that they are certainly state-of-the-art.

We did not observe apparent advantages of using higher dimensions in the case of the polynomial kernel as opposed to the results obtained by Joachims (1998b). That is,



a. Reuters b. OHSUMED

Figure. 8 Relation between C and accuracy of SVM-based categorizers. Upper plots are for polynomial kernel, and lower plots for RBF kernel.

Table 6. Accuracy of SVM classifiers (Reuters).

Kernel function	Best control parameters	Support vectors	Accuracy (dev) P/R/F	Accuracy (eval) P/R/F
poly $d = 1$	$C = 109.5$	306.6	90.84/77.54/83.67	93.48/81.18/86.89
poly $d = 2$	$C = 44.6$	308.6	91.64/77.07/83.73	93.88/80.96/86.94
poly $d = 3$	$C = 38.4$	315.1	90.79/77.60/83.68	93.67/81.29/87.04
rbf $\gamma = 0.6$	$C = 109.5$	279.7	91.42/77.18/83.70	94.47/81.51/87.51
rbf $\gamma = 0.8$	$C = 81.2$	277.2	91.37/77.18/83.67	94.76/81.29/87.51
rbf $\gamma = 1.0$	$C = 44.6$	277.6	92.23/76.60/83.69	95.14/80.46/87.18
rbf $\gamma = 2.0$	$C = 33.1$	271.2	91.77/77.18/83.84	95.06/81.18/87.57

although the C that yielded the best F-scores differed depending on d, the best F-scores we achieved did not differ a great deal. The literature often emphasizes that the advantage of the polynomial kernel is its implicit (therefore efficient) combination of features. The results seem to indicate that the combination of features is not useful for this task. However, we already showed in Section 7.3 that bi-gram features (a form of feature combination)

Table 7. Accuracy of SVM classifiers (OHSUMED).

Kernel function	Best control parameters	Support vectors	Accuracy (dev) P/R/F	Accuracy (eval) P/R/F
poly $d = 1$	$C = 84.0$	2,030.0	72.02/51.22/59.86	72.39/50.86/59.75
poly $d = 2$	$C = 43.2$	2,051.1	71.79/51.34/59.87	72.19/50.94/59.73
poly $d = 3$	$C = 27.8$	2,067.8	72.21/51.17/59.87	72.41 /50.78/59.72
rbf $\gamma = 0.6$	$C = 81.2$	2,020.7	71.43/51.59/59.91	71.95/51.23/59.85
rbf $\gamma = 0.8$	$C = 60.2$	2,023.8	71.62/51.54/59.94	72.13/51.13/59.84
rbf $\gamma = 1.0$	$C = 44.6$	2,027.8	72.22/51.33/60.01	72.62/50.82/59.80
rbf $\gamma = 2.0$	$C = 24.5$	2,062.0	72.03/51.40/59.99	71.00/51.66/59.80

improve accuracy with inequality ME models. The polynomial kernel might produce too many useless combinations that cause so severe sparseness that cannot be solved even with the large margin properties of SVMs, whereas bi-gram features produce a moderate number of combinations, which actually include useful combinations. In addition, the fact that value of the combination feature of the polynomial kernel is not TFIDF, whereas the value of a bi-gram feature is explicitly given a TFIDF value, might lead to that difference. To investigate these explanations, we evaluated an SVM categorizer using a polynomial kernel ($d = 1$) with feature vectors that explicitly contain bi-gram features represented by TFIDF values as in Section 7.3, and found that it achieved an F-score of 61.45 for the development set and 61.37 for the development set when $C = 1.2 \times 10^3$ with 3,466.2 support vectors. The achieved accuracy was close to that of the inequality ME model with bi-gram features (though still lower), demonstrating the plausibility of the above explanations.

The accuracy of the RBF kernel is better than that of the polynomial kernel, and it outperformed the Gaussian MAP ME models for both datasets and performed as well as the inequality ME models for the evaluation set of the Reuters dataset. However, it still could not outperform inequality ME models. In particular, inequality ME models are superior to SVMs with a large margin for the OHSUMED dataset.

Therefore, the most important conclusion from this experiment was that SVMs could not outperform inequality ME models, indicating that our inequality ME models are state-of-the-art for text categorization.

8. Discussion

Although the inequality ME models achieved the best performance of the models we compared, the best performance was achieved with *single* width, where a common width was used, and we could not observe any advantage in using the *bayes* width, which calculates widths depending on the unreliability of each feature. The existing regularized methods, Gaussian and exponential MAP estimation (Chen & Rosenfeld, 1999,2000; Johnson et al., 1999; Goodman, 2003,2004) can also incorporate unreliability of features through using different control parameter for each feature (e.g., σ_i). However, attempts to use a different control parameter for each feature have been rare. One exception we know of, is that by

Chen and Rosenfeld (2000), where a different σ_n is used for each level (n) of n-grams, or $\sigma_{n,1}, \sigma_{n,2}$, and $\sigma_{n,3+}$ are used for features with 1, 2, and 3 or more counts for each level of the n-grams. Chen and Rosenfeld (2000) reported that such parameter partitioning improved performance slightly over models with a single σ. However, as Chen and Rosenfeld (2000) themselves pointed out, it is unclear how to apply such partitioning to other tasks including text categorization. Another recent exception is the method presented by Goodman (2003, 2004) where control parameter a_i of the exponential prior is set so that the constraint represents Good-Turing discounting. Unfortunately, these widths based on Good-Turing discounting did not outperform the single width (Goodman, 2003, 2004).

If the single width is adequate, this might be explained as follows. By using the same width for all features, frequent features are given relatively smaller uncertainty in effect since the empirical and model expectations for frequent features are larger by default than for infrequent features. However, the discussion here is too premature to conclude that using different widths due to the unreliability of features will not be successful. It is possible that he uncertainty of $\tilde{p}(x)$, which we were not concerned with in the *bayes* width, needs to be modeled, or the Bernoulli trial assumption may be inappropriate. These matters bear further investigation.

Next, it is an open question how inequality ME models differ from the existing feature selection methods. However, inequality ME models significantly outperformed cut-off, which is closely related to DF (document frequency thresholding). As Yang and Pedersen (1997) reported, DF is competitive with other best performing feature selection methods. Therefore, we could expect that inequality ME models would also be competitive with such feature selection methods. Compared with the existing ME specific feature selection methods (Berger, Della Pietra, & Della Pietra, 1996; Della Pietra, Della Pietra, & Lafferty, 1997, Shirai et al., 1998; McCallum, 2003), the inequality ME model is more elegant than those methods in that no approximation is required in the selection process, and that feature selection is seamlessly embedded in estimation. Although our current methods for determining the constraint widths also introduce a kind of approximation about feature interaction, the interaction is exactly taken into account during estimation and the appropriate features are selected automatically, once the degree of uncertainty (i.e., the number of features) is determined by one control parameter. On the other hand, there is no such freedom in those ME specific feature selection methods, since the order of the features is determined by the approximation that abandons the exact account of feature interaction. However, performance should be compared empirically using actual tasks to clarify the advantages of each method.

Last, our justification for the advantages of inequality ME models was rather empirical, and we will need even more experiments in future. We also need to justify the advantages of inequality ME models theoretically. In the Bayesian sense, however, the inequality ME model is just a model with an exponential prior, and there is no reason to believe that the exponential prior always works best (Goodman, 2003,2004). Therefore, it would also be interesting to investigate when inequality ME models outperform other models. Such investigations might reveal why the 2-norm extension, which in effect combined an exponential prior and a Gaussian prior, seemed more robust in our experiments.

9. Conclusion and future work

We proposed inequality ME models, where the equality constraints of standard ME estimation are relaxed by using box-type inequality constraints to solve the data sparseness problem in ME estimation. We demonstrated, using two text categorization datasets, that the inequality ME models outperformed standard ME estimation, similarly motivated Gaussian MAP estimation, and state-of-the-art SVMs. The inequality ME models achieved high accuracies with a significantly small number of features thanks to the sparseness of the solution. There might be two possible directions for future work. The first is to collect more empirical evidences that prove the advantage of inequality ME models, through, for example, comparing inequality ME models with existing feature selection methods, and evaluating inequality ME models with other NLP tasks.[18] The second direction is to theoretically justify the advantages of inequality ME models, since our justification was rather empirical, or to investigate in what cases the inequality ME models outperform the other ME models.

Acknowledgments

We would like to thank Yusuke Miyao, Yoshimasa Tsuruoka, Hiroyasu Yamada, Kentaro Torisawa, and anonymous reviewers for their many helpful comments.

Appendix A. Parametric form and objective function of inequality ME model

We derive the parametric form and the dual objective function for inequality ME estimation, using the Lagrange method for convex optimization problems with linear inequality constraints.[19] To make the dual problem a maximization problem, we first rewrite the problem (Eq. 12) as:

$$\begin{aligned}
\underset{p(y|x)}{\text{minimize}} \quad & \sum_x \tilde{p}(x) \sum_y p(y \mid x) \log p(y \mid x), \\
\text{subject to} \quad & E_{\tilde{p}}[f_i] - E_p[f_i] - A_i \leq 0, \\
& E_p[f_i] - E_{\tilde{p}}[f_i] - B_i \leq 0, \\
& \sum_y p(y \mid x) - 1 = 0
\end{aligned}$$

Then, the Lagrangian is:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{p}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\mu}) = & \sum_x \tilde{p}(x) \sum_y p(y \mid x) \log p(y \mid x) \\
& + \sum_i \alpha_i (E_{\tilde{p}}[f_i] - E_p[f_i] - A_i) \\
& + \sum_i \beta_i (E_p[f_i] - E_{\tilde{p}}[f_i] - B_i) \\
& + \sum_x \tilde{p}(x) \mu_x (\sum_y p(y \mid x) - 1),
\end{aligned} \tag{33}$$

where α_i, β_i, and μ_x are the Lagrange multipliers corresponding to each constraint.

Differentiating the Lagrangian with respect to primal variables $\boldsymbol{p}$ and letting them be zero, we obtain:

$$\frac{\partial \mathcal{L}}{\partial p(y \mid x)} = \tilde{p}(x)\left\{1 + \log p(y \mid x) - \sum_i (\alpha_i - \beta_i) f_i(x, y) + \mu_i\right\} = 0.$$

Assuming $\tilde{p}(x) \neq 0$,

$$\log p(y \mid x) = -\mu_x - 1 + \sum_i (\alpha_i - \beta_i) f_i(x, y),$$

$$p(y \mid x) = \exp(-\mu_x - 1) \times \exp\left(\sum_i (\alpha_i - \beta_i) f_i(x, y)\right). \tag{34}$$

Since $\sum_y p(y \mid x) = 1$,

$$\exp(-\mu_x - 1) \sum_y \exp\left(\sum_i (\alpha_i - \beta_i) f_i(x, y)\right) = 1. \tag{35}$$

Thus,

$$\exp(-\mu_x - 1) = \left\{\sum_y \exp\left(\sum_i (\alpha_i - \beta_i) f_i(x, y)\right)\right\}^{-1} \equiv Z(x)^{-1}. \tag{36}$$

Substituting the above into Eq. 34, we obtain the parametric form for the inequality ME model as follows.

$$p_{\boldsymbol{\alpha},\boldsymbol{\beta}}(y \mid x) = \frac{1}{Z(x)} \exp\left(\sum_i (\alpha_i - \beta_i) f_i(x, y)\right). \tag{37}$$

In addition, the Lagrange method states that multipliers for inequality constraints must be greater than or equal to zero, i.e., $\alpha_i \geq 0$ and $\beta_i \geq 0$. We now have the parametric form in Eq. 15.

By substituting the parametric form into the Lagrangian (Eq. 33), we obtain the dual objective function as follows.

$$\text{Let} \quad S \equiv \sum_i (\alpha_i - \beta_i) f_i(x, y)$$

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{x,y} \tilde{p}(x) Z(x)^{-1} \exp(S)\,(S - \log Z(x))$$

$$+ \sum_i \alpha_i \left\{E_{\tilde{p}}[f_i] - \sum_{x,y} \tilde{p}(x) Z(x)^{-1} \exp(S) f_i(x, y) - A_i\right\}$$

$$
\begin{aligned}
&+ \sum_i \beta_i \left\{ \sum_{x,y} \tilde{p}(x) Z(x)^{-1} \exp(S) f_i(x, y) - E_{\tilde{p}}[f_i] - B_i \right\} \\
&= \sum_{x,y} \tilde{p}(x) Z(x)^{-1} \exp(S) \, (S - \log Z(x)) \\
&\quad - \sum_{x,y} \tilde{p}(x) Z(x)^{-1} \exp(S) \overbrace{\sum_i (\alpha_i - \beta_i) f_i(x, y)}^{S} \\
&\quad + \sum_{x,y} \overbrace{\tilde{p}(x) \tilde{p}(y \mid x)}^{\tilde{p}(x,y)} \overbrace{\sum_i (\alpha_i - \beta_i) f_i(x, y)}^{S} - \sum_i \alpha_i A_i - \sum_i \beta_i B_i \\
&= - \sum_{x,y} \tilde{p}(x) Z(x)^{-1} \exp(S) \log Z(x) + \sum_{x,y} \tilde{p}(x, y) S \\
&\quad - \sum_i \alpha_i A_i - \sum_i \beta_i B_i \\
&= - \sum_x \tilde{p}(x) Z(x)^{-1} \log Z(x) \overbrace{\sum_y \exp(S)}^{Z(x)} + \sum_{x,y} \tilde{p}(x, y) S \\
&\quad - \sum_i \alpha_i A_i - \sum_i \beta_i B_i \\
&= - \sum_x \tilde{p}(x) \log Z(x) + \sum_{x,y} \tilde{p}(x, y) S - \sum_i \alpha_i A_i - \sum_i \beta_i B_i
\end{aligned}
$$

On the other hand, the log-likelihood becomes:

$$
\begin{aligned}
LL(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \log \prod_{x,y} p_{\boldsymbol{\alpha}, \boldsymbol{\beta}}(y \mid x)^{\tilde{p}(x,y)} \\
&= \sum_{x,y} \tilde{p}(x, y) \, (S - \log Z(x)) \\
&= - \sum_x \tilde{p}(x) \left\{ \log Z(x) \sum_y \tilde{p}(y \mid x) \right\} + \sum_{x,y} \tilde{p}(x, y) S \\
&= - \sum_x \tilde{p}(x) \log Z(x) + \sum_{x,y} \tilde{p}(x, y) S
\end{aligned}
$$

Thus, there is the following relation between the dual objective function and the log-likelihood.

$$
\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = LL(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \sum_i \alpha_i A_i - \sum_i \beta_i B_i \tag{38}
$$

Notes

1. Later we show that the bounded optimization rather than the doubled parameters might be a cause of the extra cost.
2. Of course, this scenario is fictional since we should know what features are omitted using some other ways.
3. We do not claim here that the solution to the SVM always becomes sparse. However, it is true that the solution to the SVM becomes very sparse for many NLP tasks.
4. Minka (2001) also compared iterative scaling methods and gradient-based methods for logistic regression, which is very similar to ME estimation, and demonstrated the advantage of gradient-based methods.
5. Although we have only considered gradient-based methods for reasons of efficiency, extensions of GIS or IIS to support bounded parameters are also possible (Goodman, 2003,2004).
6. The term 'active' may be confusing since in the ME literature, a feature is sometimes called active when $f_i(x,y) > 0$. However, we use the term 'active' to denote that a parameter value is non-zero, following the terminology in constrained optimization. For denoting the case $f_i(x,y) > 0$, we use another term 'firing', which is used in several previous studies.
7. This is only to estimate unreliability, and is not used to calculate actual empirical expectations used in constraints.
8. See, for example, Cristianini and Shawe-Taylor (2000) for soft-margin SVMs.
9. Although C_1 and C_2 can differ for each i, we will present them here as they are common for all i.
10. We do not need to explicitly specify the bounds for δ_i and γ_i such as $\delta_i > 0$, since in the course of applying the Lagrange method, we find that $\delta_i = \frac{\alpha_i}{2C_1}$ and $\gamma_i = \frac{\beta_i}{2C_2}$, and $\delta_i, \gamma_i \geq 0$ automatically follows from $\alpha_i,\beta_i \geq 0$ and $C_1,C_2 > 0$.
11. Though actual experimental results are not presented, Goodman (2004) mentions the advantage of a double-sided model the same as ours.
12. Available from http://www.daviddlewis.com/resources/
13. For the OHSUMED dataset, we in fact extracted 30,000 documents for training, and the first 10,000 documents were used to construct the actual 9,947 training documents. The IDF value was calculated using these 30,000 documents for the OHSUMED dataset.
14. Developed by Yusuke Miyao to support various ME models such as the Feature Forest model (Miyao and Tsujii, 2002). Available at http://www-tsujii.is.s.u-tokyo.ac.jp/~yusuke/amis
15. Here, we fixed penalty constants as $C_1 = C_2 = 10^{16}$.
16. We tested at least 40 values within a range by automatically choosing the test points so that the resulting graph would become smooth. When there was a region where the curve was not smooth enough, we added experiments to smooth the region.
17. Note that these results may differ slightly from those in (Kazama and Tsujii 2003), because we have added several experimental points to each model. However, the overall arguments are the same.
18. For example, the value of the '-j' option, which handles imbalance in examples for +1 and -1 targets, and the parameters in kernel functions. However, many control parameters are sometimes a disadvantage: it is hard to conduct experiments by varying C and the coefficients of the polynomial kernel.
19. Preliminary experiments on named entity recognition can be found in Kazama (2004).
20. See, for example, Bertsekas (1999) for details on Lagrange methods.

References

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

Benson, S. J., & Moré, J. J. (2001). A limited memory variable metric method for bound constraint minimization. Technical Report ANL/MCS-P909-0901, Argonne National Laboratory.

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

Bertsekas, D. P. (1999). *Nonlinear programming,* 2nd edition. Athena Scientific.

Borthwick, A. (1999). A Maximum entropy approach to named entity recognition. Ph.D. Thesis. New York University.

Chen, S. F., & Rosenfeld, R. (1999). A gaussian prior for smoothing maximum entropy models. Technical Report CMU-CS-99-108, Carnegie Mellon University.

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

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Curran, J. R., & Clark, S. (2003). Investigating GIS and smoothing for maximum entropy taggers. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 91–98).

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

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

Fang, S.-C., Rajasekera, J. R., & Tsao, H.-S. J. (1997). *Entropy optimization and mathematical programming*. Kluwer Academic Publishers.

Goodman, J. (2003). Exponential priors for maximum entropy models. Technical Report MSR-TR-coming soon. From http://research.microsoft.com∼joshuago/ (as of September 25, 2003).

Goodman, J. (2004). Exponential priors for maximum entropy models. In *Proceedings of the HLT-NAACL 2004* (pp. 305–311).

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

Joachims, T. (1998a). Making large-scale support vector machine learning practical. In *Advances in Kernel Methods* (pp. 169–184). The MIT Press.

Joachims, T. (1998b). Text categorization with support vector machines. In *Proceedings of the 10th European Conference on Machine Learning (ECML)* (pp. 137–142).

Johnson, M., Geman, S., Canon, S., Chi, Z., & Riezler, S. (1999). Estimators for stochastic "unification-based" Grammars.

Johnson, M. & Riezler, S. (2000). Exploiting auxiliary distributions in stochastic unification-based grammars. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 154–161).

Kazama, J. (2004). Improving maximum entropy natural language processing by uncertainty-aware extensions and unsupervised learning. Ph.D. thesis, University of Tokyo.

Kazama, J., & Tsujii, J. (2003). Evaluation and extensions of maximum entropy models with inequality constraints. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 137–144).

Khudanpur, S. (1995). A method of ME estimation with relaxed constraints. In *Proceedings of Johns Hopkins University Language Modeling Workshop* (pp. 1–17).

Lau, R. (1994). Adaptive statistical language modeling. Master's Thesis. MIT.

Lewis, D. D. (1992). Feature selection and feature extraction for text categorization. In *Proceedings of Speech and Natural Language Workshop* (pp. 212–217).

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning (CoNLL)* (pp. 49–55).

McCallum, A. (2003) Efficiently inducing features of conditional random fields. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Minka, T. P. (2001). Algorithms for maximum-likelihood logistic regression. *CMU Statistics Tech Report 758*.

Miyao, Y., & Tsujii, J. (2002). Maximum entropy estimation for feature forests. In *Proceedings of Human Language Technology Conference (HLT) 2002*.

Mladenic, D., & Globelnik, M. (1998). Word sequences as features in text learning. In *Proceedings of the 17th Electrotechnical and Computer Science Conference (ERK98)*.

Newman, W (1977). Extension to the ME method. *IEEE Transactions on Information Theory, Vol. IT-23*, 89–93.

Nigam, K., John, L., & McCallum, A. (1999). Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering* (pp. 61–67).

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 79–86).

Ratnaparkhi, A. (1996). A maximum entropy model for part-of-speech tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 133–142).

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24:5.

Scott, S., & Matwin, S. (1999). Feature engineering for text classification. In *Proceedingsof the 16th International Conference on Machine Learning (ICML)* (pp. 379–388).

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. In *Proceedings of the HLT-NAACL 2003* (pp. 134–141).

Shirai, K., Inui, K., Tokunaga, T., & Tanaka, H. (1998). Saidai entoropii moderu ni yoru kakuritu moderu no parameta suitei ni yuukou na sosei no sentaku ni tuite (On the selection of useful features for estimating probabilistic models based on the maximum entropy method). In *Proceedings of 4th Annual Meeting of Natural Language Processing* (pp. 356–359). (in Japanese).

Tan, C, Wang, Y., & Lee, C (2002). The use of bigrams to enhance text categorization. *Journal Information Processing Management, 30:4*, 529–546.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer Verlag.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML)* (pp. 412–420).

Zhou, Y., Weng, F., Wu, L., & Schmidt, H. (2003). A Fast Algorithm for Feature Selection in Conditional Maximum Entropy Modeling. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 153–159).

Received October 8, 2003
Revised May 14, 2004
Accepted May 24, 2004