Maximum entropy optimization for text classification problems

Radu Ioan Boţ[1], Sorin-Mihai Grad[2], and Gert Wanka[3]

[1] Ph.D. student, Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany. Partially supported by the Gottlieb Daimler- and Karl Benz-Stiftung (under 02-48/99).
[2] Master student, Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany
[3] Professor, Faculty of Mathematics, Chemnitz University of Technology, D-09107 Chemnitz, Germany

Abstract. We present a text classification method based upon maximum entropy optimization. Having a set of documents which must be classified into some given classes, a maximum entropy optimization problem is considered. In order to solve this problem, we consider its Lagrange dual and we derive, by means of strong duality, the optimality conditions.

After solving the dual problem, we obtain, as solution of the primal problem, a distribution of probabilities representing the chances of the documents to belong to each class.

Keywords. maximum entropy optimization – Lagrange duality – optimality conditions – improved iterative scaling

1 Introduction

This paper is meant to present a rigorous way of applying the maximum entropy optimization to text classification problems, correcting the errors encountered in some other papers regarding the subject, whose authors make some compromises in order to obtain some "good-looking" results (see [3], [11]). We have a set of documents which must be classified into some given classes. A small amount of them have been a priori labelled by an expert and we have also some real-valued functions linking all the documents and the classes, called feature functions. Our goal is to obtain a distribution of probabilities of each document among the given classes.

Therefore, we impose the condition that the expected value of each feature function over the whole set of documents shall be equal with its expected value over the training sample. Using this information as constraints, we formulate the so-called maximum entropy optimization problem. Its solutions are consistent with all the constraints, but otherwise are as uniform as possible (cf. [7], [8] and [10]).

The maximum entropy optimization problem has a concave objective function and affine constraints. To solve it, we rely on the very strong results of the theory of duality. This gives us the possibility to formulate for an optimization problem its dual. Moreover, by means of strong duality the optimality conditions can be derived (cf. [12]).

For our maximum entropy optimization problem we develop, here, the Lagrange dual. As a consequence of the optimality conditions, we write the solutions of the primal problem as functions of solutions of the dual problem. The last ones are determined using the so-called iterative scaling algorithm developed from the one introduced by Darroch and Ratcliff (cf. [4], [11]).

Finally, by the use of the solutions of the dual, we find the desired distribution of probabilities.

2 The formulation of the problem

Let us consider a set of documents $\mathcal{D}$ and the set of classes $\mathcal{C}$ where they must be classified into. There is also a given subset of $\mathcal{D}$, denoted $\mathcal{D}'$, whose elements have been labelled by an expert as to belong to a certain class from $\mathcal{C}$. To have information about all the classes, we need to consider that each class contains at least an element from $\mathcal{D}'$. One may notice that between the sets $\mathcal{C}$ and $\mathcal{D}'$, it must hold $|\mathcal{C}| \leq |\mathcal{D}'|$, where $|\mathcal{C}|$ is the cardinal of the set $\mathcal{C}$ and $|\mathcal{D}'|$ is the cardinal of the set $\mathcal{D}'$. The set of pairs $(d', c(d')), d' \in \mathcal{D}'$, obtained above, is called the training data and $c(d') \in \mathcal{C}$ denotes the class which is assigned to d' by the expert.

The labelled training data set is used to derive a set of constraints for the model that characterize the class-specific expectations for the distribution. Constraints are represented as expected values of so-called features functions, which may be any positive real-valued functions defined over $\mathcal{D} \times \mathcal{C}$. Let us denote by $f_i, i \in I$, the feature functions for the problem of text classification, which we treat here.

As an example, we will present the set of feature functions considered in [11] for the same problem of text classification. Denoting by $\mathcal{W}$ the set of the words which appear in the whole family of documents $\mathcal{D}$, the set I is defined by

$$I = \mathcal{W} \times \mathcal{C}.$$

For each word-class combination $(w, c') \in \mathcal{W} \times \mathcal{C}$, one can consider the feature function $f_{w,c'} : \mathcal{W} \times \mathcal{C} \to \mathbb{R}$,

$$f_{w,c'}(d, c) = \begin{cases} 0, & c \neq c', \\ \frac{N(d,w)}{N(d)}, & \text{otherwise,} \end{cases}$$

where $N(d, w)$ is the number of times word w occurs in document d, $N(d)$ is the number of words in d, and c, c' are classes in $\mathcal{C}$.

Other ways to consider feature functions can be found in [1] and [9].

Using the information given by this training data and the feature functions, we want to obtain the distribution of probabilities of each document $d \in \mathcal{D}$ among the given classes. The way we are to use for this is quite heuristical (cf. [1], [9]), consisting in generalizing some facts that hold for the training sample to the whole set of documents. The expected value of each feature function over all documents and classes will be forced to equal its expected value over the training sample

$$\tilde{E}(f_i) = E(f_i), \forall i \in I. \tag{1}$$

The expected value of each feature function $f_i, i \in I$, regarding the training sample comes from the following formula

$$\tilde{E}(f_i) = \sum_{d' \in \mathcal{D}'} \sum_{c \in \mathcal{C}} p(d', c) f_i(d', c), i \in I, \tag{2}$$

where $p(d', c)$ denotes the joint probability of c and d'. But this joint probability can be decomposed as

$$p(d', c) = p(d')p(c|d'),$$

with $p(d')$ being the probability of the document d' to be chosen from the training data and $p(c|d')$ the conditional probability of the class c with respect to the document d'.

The probability of the document d' to be chosen from the training data is

$$p(d') = \frac{1}{|\mathcal{D}'|}, \text{ for } d' \in \mathcal{D}'.$$

On the other hand, as we know that each document from the training data has been a priori labelled, it is clear that

$$p(c|d') = \begin{cases} 1, & \text{if } c = c(d'), \\ 0, & \text{if } c \neq c(d'), \end{cases}$$

for every $c \in \mathcal{C}$ and $d' \in \mathcal{D}'$.

By (2), we have then

$$\tilde{E}(f_i) = \frac{1}{|\mathcal{D}'|} \sum_{d' \in \mathcal{D}'} f_i(d', c(d')), i \in I. \tag{3}$$

The expected value of f_i regarding the whole set $\mathcal{D} \times \mathcal{C}$ is

$$E(f_i) = \sum_{d \in \mathcal{D}} \sum_{c \in \mathcal{C}} p(d, c) f_i(d, c), i \in I.$$

As for the training set, we have

$$p(d, c) = p(d)p(c|d),$$

with $p(d) = \frac{1}{|\mathcal{D}|}$ being the probability to choose the document d from $\mathcal{D}$ and $p(c|d)$ the conditional probability of the class c with respect to the document d. It follows

$$E(f_i) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sum_{c \in \mathcal{C}} p(c|d) f_i(d, c), i \in I. \tag{4}$$

For each feature $f_i, i \in I$, we will constrain now the model to have the same expected value for it over the whole set of documents as the one obtained from the training set. From (1), (3) and (4), we obtain

$$\frac{1}{|\mathcal{D}'|} \sum_{d' \in \mathcal{D}'} f_i(d', c(d')) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sum_{c \in \mathcal{C}} p(c|d) f_i(d, c), i \in I. \tag{5}$$

Moreover, from the basic properties of the probability distributions, it holds

$$p(c|d) \geq 0, \forall c \in \mathcal{C}, \forall d \in \mathcal{D}, \tag{6}$$

and

$$\sum_{c \in \mathcal{C}} p(c|d) = 1, \forall d \in \mathcal{D}. \tag{7}$$

The problem that we have to solve now is to find a probability distribution which fulfills the constraints (5), (6) and (7). Therefore, we will use a technique which bases on theory of maximum entropy (cf. [7], [8] and [10]). The over-riding principle in maximum entropy is that when nothing is known, the distribution of probabilities should be as uniform as possible.

But, that is exactly what results by solving the following so-called maximum entropy optimization problem

$$(P) \quad \sup \{H(p)\},$$

subject to

$$\frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum_{d' \in \mathcal{D}'} f_i(d', c(d')) = \sum_{d \in \mathcal{D}} \sum_{c \in \mathcal{C}} p(c|d) f_i(d, c), \forall i \in I,$$

$$\sum_{c \in \mathcal{C}} p(c|d) = 1, \forall d \in \mathcal{D},$$

and

$$p(c|d) \geq 0, \forall c \in \mathcal{C}, \forall d \in \mathcal{D}.$$

Here, $H : \mathbb{R}^{|C|\cdot|D|} \to \overline{\mathbb{R}}$ is the so-called entropy function and it is defined, for $p = (p(c|d))_{c\in\mathcal{C},d\in\mathcal{D}}$, by

$$H(p) = \begin{cases} -\sum\limits_{d\in\mathcal{D}}\sum\limits_{c\in\mathcal{C}} p(c|d)\ln p(c|d), \text{ if } p(c|d) \geq 0, \forall c \in \mathcal{C}, \forall d \in \mathcal{D}, \\ -\infty, \qquad\qquad\qquad\qquad \text{otherwise.} \end{cases}$$

It is obvious that H is a concave function.

3 Duality for the maximum entropy optimization problem

The goal of this chapter is to formulate a dual problem to the maximum entropy optimization problem

$$(P) \quad \sup\left\{-\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} p(c|d)\ln p(c|d)\right\},$$

subject to

$$\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{d'\in\mathcal{D}'} f_i(d', c(d')) = \sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} p(c|d)f_i(d, c), \forall i \in I,$$

$$\sum_{c\in\mathcal{C}} p(c|d) = 1, \forall d \in \mathcal{D},$$

$$p(c|d) \geq 0, \forall c \in \mathcal{C}, \forall d \in \mathcal{D},$$

and to derive, by means of strong duality, the optimality conditions for (P) and its dual.

Therefore, we will consider the following general primal optimization problem

$$(P_g) \quad \inf_{\substack{g(x)=0,\\ x\in X}} f(x),$$

where $X \subseteq \mathbb{R}^n$ is a non-empty convex set, $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is a convex function and $g : \mathbb{R}^n \to \mathbb{R}^m$, $g(x) = (g_1(x), ..., g_m(x))^T$, such that $g_i : \mathbb{R}^n \to \mathbb{R}$, $i = 1, ..., m$, are affine functions.

The Lagrange dual problem to (P_g) is given by

$$(D_g) \quad \sup_{\lambda\in\mathbb{R}^m}\inf_{x\in X}\left\{f(x) + \lambda^T g(x)\right\}.$$

It is trivial to show that the weak duality between (P_g) and (D_g) always holds, i.e. $\inf(P_g) \geq \sup(D_g)$. For the strong duality, $\inf(P_g) = \sup(D_g)$, we need to consider the fulfillment of a constraint qualification. The functions $g_i, i = 1, ..., m$, being affine, we can consider in this case the following Slater constraint qualification (cf. [6])

$$(SCQ_g): \quad \text{there exists } x' \in rintX, \text{ such that } g(x') = 0,$$

where $rintX$ is the relative interior of the set X.

Let us present now the strong duality theorem and formulate the optimality conditions for (P_g) and (D_g) (cf. [6], [12]).

Theorem 1. *Let us assume that* $\inf(P_g) > -\infty$ *and that the constraint qualification* (SCQ_g) *is fulfilled. Then the dual problem has a solution and between* (P_g) *and* (D_g) *strong duality holds,*

$$\inf(P_g) = \max(D_g).$$

Theorem 2. *Let the constraint qualification (SCQ_g) be fulfilled. Then $\bar{x}$ is a solution to (P_g) if and only if $\bar{x}$ is feasible to (P_g) and there exists $\bar{\lambda} \in \mathbb{R}^m$, such that the following conditions are satisfied*

$$(i)\ \inf_{x\in X}\left[f(x)+\sum_{i=1}^{m}\bar{\lambda}_i g_i(x)\right]=f(\bar{x}),$$
$$(ii)\ \sum_{i=1}^{m}\bar{\lambda}_i g_i(\bar{x})=0.$$

In order to study the duality for the initial problem (P), we need to consider another optimization problem

$$(P')\quad \inf\left\{\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)\ln p(c|d)\right\},$$

subject to

$$\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{d'\in\mathcal{D}'}f_i(d',c(d'))=\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)f_i(d,c),\forall i\in I,$$
$$\sum_{c\in\mathcal{C}}p(c|d)=1,\forall d\in\mathcal{D},$$
$$p\in X,$$

with $X=\{p=(p(c|d))_{c\in\mathcal{C},d\in\mathcal{D}}: p(c|d)\geq 0,\forall c\in\mathcal{C},\forall d\in\mathcal{D}\}$. The problem (P') fits in the scheme already presented and has the same solutions as (P), so that it holds

$$\sup(P)=-\inf(P').$$

According to the general case, its dual problem is

$$(D')\quad \sup_{\substack{\lambda_i\in\mathbb{R},i\in I,\\ \lambda_d\in\mathbb{R},d\in\mathcal{D}}}\ \inf_{\substack{p(c|d)\geq 0,\\ (d,c)\in\mathcal{D}\times\mathcal{C}}}\left[\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)\ln p(c|d)+\sum_{d\in\mathcal{D}}\lambda_d\left(\sum_{c\in\mathcal{C}}p(c|d)-1\right)\right.$$
$$\left.+\sum_{i\in I}\lambda_i\left(\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{d'\in\mathcal{D}'}f_i(d',c(d'))-\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)f_i(d,c)\right)\right].$$

Like above, we can find another problem, (D''), which has the same solutions as (D') so that $\sup(D')=-\inf(D'')$,

$$(D'')\quad \inf_{\substack{\lambda_i\in\mathbb{R},i\in I,\\ \lambda_d\in\mathbb{R},d\in\mathcal{D}}}\ \sup_{\substack{p(c|d)\geq 0,\\ (d,c)\in\mathcal{D}\times\mathcal{C}}}\left[-\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)\ln p(c|d)-\sum_{d\in\mathcal{D}}\lambda_d\left(\sum_{c\in\mathcal{C}}p(c|d)-1\right)\right.$$
$$\left.-\sum_{i\in I}\lambda_i\left(\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{d'\in\mathcal{D}'}f_i(d',c(d'))-\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)f_i(d,c)\right)\right].$$

The problem (D'') can be rewritten as

$$(D'')\quad \inf_{\substack{\lambda_i\in\mathbb{R},i\in I,\\ \lambda_d\in\mathbb{R},d\in\mathcal{D}}}\left\{\sum_{d\in\mathcal{D}}\lambda_d-\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\lambda_i\sum_{d'\in\mathcal{D}'}f_i(d',c(d'))+\right.$$
$$\left.\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}\sup_{p(c|d)\geq 0}\left[-p(c|d)\ln p(c|d)-\lambda_d p(c|d)+\sum_{i\in I}\lambda_i p(c|d)f_i(d,c)\right]\right\}.$$

To calculate the suprema which appear in the above formula, we shall consider the function

$$u:\mathbb{R}_+\to\mathbb{R},u(x)=-x\ln x-\lambda_d x+x\sum_{i\in I}\lambda_i f_i(d,c),x\in\mathbb{R}_+,$$

for some fixed $(d,c) \in \mathcal{D} \times \mathcal{C}$.

Its derivative is

$$u'(x) = -\ln x - 1 - \lambda_d + \sum_{i \in I} \lambda_i f_i(d,c),$$

and it holds

$$u'(x) = 0 \Leftrightarrow \ln x = -\lambda_d - 1 + \sum_{i \in I} \lambda_i f_i(d,c) \Leftrightarrow x = e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)} > 0.$$

The function u being concave, it follows that at $x = e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)}$ it attains its maximal value. So

$$\begin{aligned}
\max_{x \geq 0} u(x) &= u\left(e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)}\right) \\
&= -e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)} \left(\ln e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)} + \lambda_d - \sum_{i \in I} \lambda_i f_i(d,c)\right) \\
&= -e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)} \left(-\lambda_d - 1 + \sum_{i \in I} \lambda_i f_i(d,c) + \lambda_d - \sum_{i \in I} \lambda_i f_i(d,c)\right) \\
&= e^{-\lambda_d - 1 + \sum\limits_{i \in I} \lambda_i f_i(d,c)}.
\end{aligned}$$

Our dual problem becomes then

$$\begin{aligned}
(D'') \quad \inf_{\substack{\lambda_i \in \mathbb{R}, i \in I, \\ \lambda_d \in \mathbb{R}, d \in \mathcal{D}}} \Bigg\{ & \sum_{d \in \mathcal{D}} \lambda_d - \frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum_{i \in I} \lambda_i \sum_{d' \in \mathcal{D}'} f_i(d', c(d')) \\
& + \sum_{d \in \mathcal{D}} \sum_{c \in \mathcal{C}} e^{\sum\limits_{i \in I} \lambda_i f_i(d,c) - \lambda_d - 1} \Bigg\}.
\end{aligned}$$

In the next part of the section we will make some assertions concerning the duality between (P) and (D''). For this, we will apply the results formulated in the general case for the problems (P_g) and (D_g). Let us write, first, the Slater constraint qualification for the problems (P) and, respectively, (P')

(SCQ): there exists $p' = (p'(c|d))_{c \in \mathcal{C}, d \in \mathcal{D}}$, such that

$$\begin{cases}
p'(c|d) > 0, \forall c \in \mathcal{C}, \forall d \in \mathcal{D}, \\
\frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum\limits_{d' \in \mathcal{D}'} f_i(d', c(d')) = \sum\limits_{d \in \mathcal{D}} \sum\limits_{c \in \mathcal{C}} p'(c|d) f_i(d,c), \forall i \in I, \\
\sum\limits_{c \in \mathcal{C}} p'(c|d) = 1, \forall d \in \mathcal{D}.
\end{cases}$$

By this, we can state now the desired strong duality theorem and the optimality conditions for (P), by applying Theorem 1 and Theorem 2, respectively.

Theorem 3. *Consider the problem (P') introduced above and let (SCQ) be fulfilled. Its dual problem has then a solution and between (P') and (D') strong duality holds,*

$$\sup(P) = -\inf(P') = -\max(D') = \min(D'').$$

Theorem 4. *Let us assume that the constraint qualification (SCQ) is fulfilled. Then $\bar{p} = ((\bar{p}(c|d))_{c \in \mathcal{C}, d \in \mathcal{D}}$ is a solution to (P) if and only if $\bar{p}$ is feasible to (P) and there exist $\bar{\lambda}_i \in \mathbb{R}, i \in I$, and $\bar{\lambda}_d \in \mathbb{R}, d \in \mathcal{D}$, such that the following conditions are*

satisfied

$$(i)\quad \inf_{\substack{p(c|d)\geq 0,\\ d\in\mathcal{D},c\in\mathcal{C}}} \Bigg[\sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} p(c|d)\ln p(c|d) + \sum_{d\in\mathcal{D}} \bar{\lambda}_d \left(\sum_{c\in\mathcal{C}} p(c|d) - 1\right)$$

$$+ \sum_{i\in I} \bar{\lambda}_i \left(\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{d'\in\mathcal{D}'} f_i(d', c(d')) - \sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} p(c|d) f_i(d,c)\right)\Bigg]$$

$$= \sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} \bar{p}(c|d)\ln\bar{p}(c|d)$$

$$(ii)\quad 0 = \sum_{i\in I} \bar{\lambda}_i \left(\frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{d'\in\mathcal{D}'} f_i(d', c(d')) - \sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}} \bar{p}(c|d) f_i(d,c)\right)$$

$$+ \sum_{d\in\mathcal{D}} \bar{\lambda}_d \left(\sum_{c\in\mathcal{C}} \bar{p}(c|d) - 1\right).$$

Remark 1. Let us point out that all the functions involved in the formulation of the primal problem are differentiable. This implies that the equality (i) in Theorem 4 can be, equivalently, written as

$$\ln p(c|d) + 1 + \bar{\lambda}_d - \sum_{i\in I} \bar{\lambda}_i f_i(d,c) = 0, \forall d\in\mathcal{D}, \forall c\in\mathcal{C},$$

or

$$p(c|d) = \frac{e^{\sum\limits_{i\in I}\bar{\lambda}_i f_i(d,c)}}{e^{\bar{\lambda}_d+1}}, \forall d\in\mathcal{D}, \forall c\in\mathcal{C}. \tag{8}$$

Getting now back to the problem (D''), one may observe that it can be decomposed into

$$(D'')\quad \inf_{\substack{\lambda_i\in\mathbb{R},\\ i\in I}} \Bigg\{\sum_{d\in\mathcal{D}} \inf_{\lambda_d\in\mathbb{R}} \left[\sum_{c\in\mathcal{C}} e^{\sum\limits_{i\in I}\lambda_i f_i(d,c) - \lambda_d - 1} + \lambda_d\right]$$

$$- \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\lambda_i \sum_{d'\in\mathcal{D}'} f_i(d', c(d'))\Bigg\}.$$

We can calculate the infima inside the parentheses, using another auxiliary function

$$v:\mathbb{R}\to\mathbb{R}, v(x) = e^{-x-1}a + x, x\in\mathbb{R}, a>0.$$

It is convex and derivable, its derivative

$$v'(x) = 1 - ae^{-x-1}$$

fulfilling

$$v'(x) = 0 \Leftrightarrow e^{-x-1} = \frac{1}{a} \Leftrightarrow x = \ln a - 1.$$

So, v's minimum is attained at $\ln a - 1$, being

$$v(\ln a - 1) = \ln a.$$

Taking $a = \sum\limits_{c\in\mathcal{C}} e^{\sum\limits_{i\in I}\lambda_i f_i(d,c)} > 0$, the dual problem turns into

$$(D)\quad \inf_{\substack{\lambda_i\in\mathbb{R},\\ i\in I}} \left[\sum_{d\in\mathcal{D}} \ln \sum_{c\in\mathcal{C}} e^{\sum\limits_{i\in I}\lambda_i f_i(d,c)} - \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\lambda_i \sum_{d'\in\mathcal{D}'} f_i(d', c(d'))\right]$$

and, obviously, we have

$$\inf(D'') = \inf(D).$$

In fact, we have proved the following assertion concerning the solutions of the problems (D) and (D'').

Theorem 5. *The following equivalence holds*

$$((\bar{\lambda}_d)_{d\in\mathcal{D}}, (\bar{\lambda}_i)_{i\in I}) \text{ is a solution to } (D'') \Leftrightarrow \begin{cases} (\bar{\lambda}_i)_{i\in I}) \text{ is a solution to } (D) \text{ and} \\ \bar{\lambda}_d = \ln \sum\limits_{c\in\mathcal{C}} e^{\sum\limits_{i\in I} \bar{\lambda}_i f_i(d,c)} - 1, \forall d \in \mathcal{D}. \end{cases}$$

Remark 2. By Remark 1 and Theorem 5 it follows that, in order to find a solution of the problem (P), it is enough to solve the dual problem (D). Getting $(\bar{\lambda}_i)_{i\in I}$, solution to (D), we obtain, for each $d \in \mathcal{D}$,

$$\bar{\lambda}_d = \ln \sum_{c\in\mathcal{C}} e^{\sum\limits_{i\in I} \bar{\lambda}_i f_i(d,c)} - 1$$

and, by (8),

$$p(c|d) = \frac{e^{\sum\limits_{i\in I} \bar{\lambda}_i f_i(d,c)}}{e^{\bar{\lambda}_d + 1}} = \frac{e^{\sum\limits_{i\in I} \bar{\lambda}_i f_i(d,c)}}{\sum\limits_{c\in\mathcal{C}} e^{\sum\limits_{i\in I} \bar{\lambda}_i f_i(d,c)}}, \forall (d,c) \in \mathcal{D}\times\mathcal{C}. \tag{9}$$

4 Solving the dual problem

In this section, we will outline the derivation of an algorithm for finding a solution of the dual problem (D). The algorithm is called improved iterative scaling and other variants of it have been described by different authors in connection with maximum entropy optimization problems (cf. [3] and [11]).

First, let us introduce the function $l : \mathbb{R}^{|I|} \to \mathbb{R}$, defined by

$$l(\lambda) = \sum_{i\in I} \lambda_i \left(\frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum_{d'\in\mathcal{D}'} f_i(d', c(d')) \right) - \sum_{d\in\mathcal{D}} \ln \sum_{c\in\mathcal{C}} e^{\sum\limits_{i\in I} \lambda_i f_i(d,c)},$$

for $\lambda = (\lambda_i)_{i\in I}$.

Considering the optimization problem

$$(P_l) \quad \max_{\lambda\in\mathbb{R}^{|I|}} l(\lambda),$$

it is obvious that

$$\min(D) = -\max(P_l),$$

and that the sets of the solutions of the two problems coincide. So, in order to obtain the desired results, it is enough to solve (P_l).

Let us calculate now, for $\lambda = (\lambda_i)_{i\in I}, \delta = (\delta_i)_{i\in I} \in \mathbb{R}^{|I|}$, the expression $\Delta l := l(\lambda+\delta) - l(\lambda)$. It holds

$$\begin{aligned}\Delta l &= \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}(\lambda_i+\delta_i)f_i(d',c(d')) - \sum_{d\in\mathcal{D}}\ln\sum_{c\in\mathcal{C}}e^{\sum_{i\in I}(\lambda_i+\delta_i)f_i(d,c)}\\ &- \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\lambda_i f_i(d',c(d')) + \sum_{d\in\mathcal{D}}\ln\sum_{c\in\mathcal{C}}e^{\sum_{i\in I}\lambda_i f_i(d,c)}\\ &= \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\delta_i f_i(d',c(d')) - \sum_{d\in\mathcal{D}}\ln\frac{\sum_{c\in\mathcal{C}}e^{\sum_{i\in I}(\lambda_i+\delta_i)f_i(d,c)}}{\sum_{c\in\mathcal{C}}e^{\sum_{i\in I}\lambda_i f_i(d,c)}}.\end{aligned}$$

As it is known that

$$-\ln(x) \geq 1-x, \forall x\in\mathbb{R}_+,$$

we have

$$\begin{aligned}\Delta l &\geq \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\delta_i f_i(d',c(d')) + \sum_{d\in\mathcal{D}}\left(1-\frac{\sum_{c\in\mathcal{C}}e^{\sum_{i\in I}(\lambda_i+\delta_i)f_i(d,c)}}{\sum_{c\in\mathcal{C}}e^{\sum_{i\in I}\lambda_i f_i(d,c)}}\right)\\ &= \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\delta_i f_i(d',c(d')) + \sum_{d\in\mathcal{D}}\left(1-\sum_{c\in\mathcal{C}}p(c|d)e^{\sum_{i\in I}\delta_i f_i(d,c)}\right).\end{aligned}$$

Denoting

$$f^{\#}(d,c) = \sum_{i\in I} f_i(d,c),$$

we get

$$\Delta l \geq \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\delta_i f_i(d',c(d')) + \sum_{d\in\mathcal{D}}\left(1-\sum_{c\in\mathcal{C}}p(c|d)e^{f^{\#}(d,c)\sum_{i\in I}\delta_i\frac{f_i(d,c)}{f^{\#}(d,c)}}\right).$$

As the exponential function is convex, applying Jensen's inequality

$$e^{f^{\#}(d,c)\sum_{i\in I}\delta_i\frac{f_i(d,c)}{f^{\#}(d,c)}} \leq \sum_{i\in I}\frac{f_i(d,c)}{f^{\#}(d,c)}e^{f^{\#}(d,c)\delta_i},$$

there follows

$$\Delta l \geq \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\delta_i f_i(d',c(d')) - \sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)\sum_{i\in I}\frac{f_i(d,c)}{f^{\#}(d,c)}e^{f^{\#}(d,c)\delta_i} + |\mathcal{D}|.$$

Let be now $\mathcal{B}:\mathbb{R}^{|I|}\to\mathbb{R}$, the following function

$$\mathcal{B}(\delta) = \frac{|\mathcal{D}|}{|\mathcal{D}'|}\sum_{i\in I}\sum_{d'\in\mathcal{D}'}\delta_i f_i(d',c(d')) - \sum_{d\in\mathcal{D}}\sum_{c\in\mathcal{C}}p(c|d)\sum_{i\in I}\frac{f_i(d,c)}{f^{\#}(d,c)}e^{f^{\#}(d,c)\delta_i} + |\mathcal{D}|,$$

for $\delta = (\delta_i)_{i\in I}$.

We can guarantee an increase of the value of the function l if we can find a δ such that $B(\delta)$ is positive. B is a concave function since its first term is a linear function, the second contains a sum of concave functions and the third is a constant. Moreover,

$\mathcal{B}$ is a differentiable function. So, to find the best δ, we need to differentiate $B(\delta)$ with respect to the change in each parameter δ_i, $i \in I$, and to set

$$\frac{\partial \mathcal{B}}{\partial \delta_i} = 0, \forall i \in I.$$

We get

$$\frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum_{d' \in \mathcal{D}'} f_i(d', c(d')) = \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} p(c|d) f_i(d, c) e^{f^{\#}(d,c)\delta_i}, \forall i \in I.$$

Solving these equations we obtain the values of $\delta_i, i \in I$. In the next section we shall present the complete algorithm to determine the maximum of the function l.

Remark 3. We have to mention here that in [3] and [11] the function l has been identified with the so-called maximum likelihood, whose formula is considered

$$L(\lambda) = \ln \prod_{d' \in \mathcal{D}'} p(c(d')|d').$$

This is possible only if one considers the sets $\mathcal{D}$ and $\mathcal{D}'$ identical. In this case, we have

$$\begin{aligned} l(\lambda) &= \sum_{i \in I} \lambda_i \left(\frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum_{d' \in \mathcal{D}'} f_i(d', c(d')) \right) - \sum_{d \in \mathcal{D}} \ln \sum_{c \in \mathcal{C}} e^{\sum_{i \in I} \lambda_i f_i(d,c)} \\ &= \sum_{d' \in \mathcal{D}'} \left[\sum_{i \in I} \lambda_i f_i(d', c(d')) - \ln \sum_{c \in \mathcal{C}} e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))} \right] \\ &= \sum_{d' \in \mathcal{D}'} \left[\ln e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))} - \ln \sum_{c \in \mathcal{C}} e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))} \right] \\ &= \sum_{d' \in \mathcal{D}'} \left[\ln \frac{e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))}}{\sum_{c \in \mathcal{C}} e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))}} \right] \end{aligned}$$

Finally, using the relations given in (9), the function l turns out to be in this case identical to the maximum likelihood function

$$l(\lambda) = \sum_{d' \in \mathcal{D}'} \left[\ln \frac{e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))}}{\sum_{c \in \mathcal{C}} e^{\sum_{i \in I} \lambda_i f_i(d', c(d'))}} \right] = \sum_{d' \in \mathcal{D}'} \ln p(c(d')|d') = \ln \prod_{d' \in \mathcal{D}'} p(c(d')|d').$$

We can conclude that the results obtained in [3] and [11] do not refer to the unclassified documents using the information given by that expert regarding the training sample, being just distributions of the same a priori labelled documents among all the classes. We consider that this compromise is not useful in our problem, as we have proved before that the algorithm works also without it.

5 An algorithm for solving the maximum entropy optimization problem

In this section we will present, by the use of the results obtained in the previous sections, an algorithm for solving the dual of the maximum entropy optimization

problem. Assuming that the Slater constraint qualification (SCQ) is fulfilled the solutions of the primal problem arise by calling (9). This is a generalization of the algorithm introduced by Darroch and Ratcliff in [4].

Inputs: A collection $\mathcal{D}$ of documents, a subset of it $\mathcal{D}'$ of labelled documents, a set of classes $\mathcal{C}$ and a set of feature functions $f_i, i \in I$, connecting the documents and the classes. Let $\varepsilon > 0$ be the admitted error of the iterative process.

Step 1: Set the constraints. For every feature $f_i, i \in I$, estimate its expected value over the set of the documents and the set of classes.

Step 2: Set the initial values $\lambda_i = 0, i \in I$.

Step 3:

- Using the equalities in (9), calculate with the current parameters $(\lambda_i)_{i \in I}$ the values for $p(c|d), (d, c) \in \mathcal{D} \times \mathcal{C}$.
- for each $i \in I$:
  - find δ_i, a solution of the equation

$$\frac{|\mathcal{D}|}{|\mathcal{D}'|} \sum_{d' \in \mathcal{D}'} f_i(d', c(d')) = \sum_{c \in \mathcal{C}} \sum_{d \in \mathcal{D}} p(c|d) f_i(d, c) e^{f^{\#}(d,c)\delta_i}.$$

  - set $\lambda_i = \lambda_i + \delta_i$.

Step 4: If there exists an $i \in I$, such that $|\delta_i| > \varepsilon$, then go to **Step 3**.

Remark 4. (a) By setting $\lambda_i = 0, \forall i \in I$, the initial values for the probability distributions are

$$p(c|d) = \frac{1}{|\mathcal{C}|}, c \in \mathcal{C}, d \in \mathcal{D}.$$

(b) In the original algorithm Darroch and Ratcliff assumed in [4] that $f^{\#}(d, c)$ is constant. Denoting its value by M, one gets then

$$\delta_i = \frac{1}{M} ln \left(\frac{|\mathcal{D}|}{|\mathcal{D}'|} \frac{\sum\limits_{d' \in \mathcal{D}'} f_i(d', c(d'))}{\sum\limits_{c \in \mathcal{C}} \sum\limits_{d \in \mathcal{D}} p(c|d) f_i(d, c)} \right), i \in I.$$

(c) A more detailed discussion regarding the iterative scaling algorithm, including a proof of its convergence, can be found in [2], [4], [5] and [11].

Having obtained $\lambda_i, i \in I$, returned by the algorithm, we can determine by (9) the solutions of the primal problem, i.e. the probability distributions of each document among the given classes.

To assign each document with a certain class, one can consider more criteria, such as to choose the class whose probability is the highest, or to establish a minimal value of probability and to label the documents as belonging to all the classes that fit, and if neither does, to create a supplementer class for this document. But these criteria debates surpass the purpose of the present paper.

Acknowledgements. The authors would like to thank anonymous referee for his valuable and helpful suggestions.

References

1. Baeza-Yates, R., Ribeiro-Neto, B. (1999) Modern Information Retrieval. Addison-Wesley Verlag
2. Bapat, R.B., Raghavan, T.E.S. (1997) Nonnegative Matrices and Applications. Cambridge University Press, Cambridge

3. Berger, A.L., Della Pietra, S.A., Della Pietra, V.J. (1996) A Maximum Entropy Approach to Natural Language Processing. Comput. Ling. **22, 1**, 1–36
4. Darroch, J.N., Ratcliff, D. (1972) Genelarized iterative scaling for log-linear models. Ann. Math. Stat. **43**, 1470–1480
5. Della Pietra, S., Della Pietra, V., Lafferty, J. (1997) Inducing Features of Random Fields. IEEE Trans. Pattern Anal. Mach. Intell. **19, 4**, 1–13
6. Elster, K.H., Reinhart, R., Schäuble, M., Donath, G. (1977) Einführung in die nichtlineare Optimierung. BSB B.G. Teubner Verlagsgesellschaft, Leipzig
7. Fang, S.C., Rajasekera, J.R., Tsao, H.-S.J. (1997) Entropy Optimization and Mathematical Programming. Kluwer Academic Publishers, Boston
8. Guiaşu, S., Shenitzer, A. (1985) The Principle of Maximum Entropy. The Math. Intell. **7, 1**, 42–48
9. Jurafsky, D., Martin, J.H. (2000) Speech and Language Processing. Prentice-Hall Inc.
10. Kapur, J.N., Kesavan, H.K. (1992) Entropy Optimization Principles with Applications. Academic Press Inc., San Diego
11. Nigam, K., Lafferty, J., McCallum, A. (1999) Using Maximum Entropy for Text Classification. Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering, 61–67
12. Wanka, G., Boţ, R.I. (2002) On the Relations Between Different Dual Problems in Convex Mathematical Programming. In Chamoni, P., Leisten, R., Martin, A., Minnemann, J., Stadtler, H., (eds.), Operations Research Proceedings 2001, Springer, Heidelberg, 255–262