DISCRIMINATIVE TRAINING VIA LINEAR PROGRAMMING *

Kishore A. Papineni
IBM T. J. Watson Research Center
Yorktown Heights, NY 10598, USA
Email: kishore@watson.ibm.com

ABSTRACT

This paper presents a linear programming approach to discriminative training. We first define a measure of discrimination of an arbitrary conditional probability model on a set of labeled training data. We consider maximizing discrimination on a parametric family of exponential models that arises naturally in the maximum entropy framework. We show that this optimization problem is globally convex in R^n, and is moreover piece-wise linear on R^n. We propose a solution that involves solving a series of linear programming problems. We provide a characterization of global optimizers. We compare this framework with those of minimum classification error and maximum entropy.

1. INTRODUCTION

Consider conditional probability distributions $P(f|h)$ where h is *history* and f is *future*. Our goal is to predict the future using the model P, given the history. In a classification problem, history is the observation vector and future is the label of a class. As in a classification problem, consider the case when there is only a finite set of futures $\mathcal{F}$, fixed *a priori*.

Let the model's best guess of future be

$$\arg\max_{f\in\mathcal{F}} P(f|h).$$

In the classification problem, this is the Bayes decision rule or the *maximum a posteriori* (MAP) decision rule, and leads to minimum error rate classification when used with the true a posteriori probability distribution on the underlying variables. However, true distribution is not available to us and can only be estimated from a set of labeled training samples.

Suppose we are given a collection of training pairs (h_i, f_i), $i = 1, \ldots, T$. Treating training data as truth, we ideally want a model P such that for each i,

$$P(f_i|h_i) = \max_{f\in\mathcal{F}} P(f|h_i).$$

This will rarely be the case for a model. We want to assign a measure of goodness to the model that penalizes wrong guesses; the penalty shall be proportional to how far away the guess is from truth in terms of probabilities.

2. OBJECTIVE IS DISCRIMINATION

We consider two measures of *discrimination* of the model P. One measure is defined by

$$D_1(P) := \sum_{i=1}^{T} \log \frac{P(f_i|h_i)}{\max_{f\in\mathcal{F}} P(f|h_i)}. \qquad (1)$$

The second is defined by

$$D_2(P) := \sum_{i=1}^{T} \log \frac{P(f_i|h_i)}{\max_{f\in\mathcal{F}-\{f_i\}} P(f|h_i)}. \qquad (2)$$

Clearly,

$$D_2(P) \geq D_1(P) \geq \sum_{i=1}^{T} \log P(f_i|h_i) =: L(P)$$

where $L(P)$ is the likelihood of the training data according to the model P. Also, $D_1(P) \leq 0$ for any model P.

The motivation behind these definitions is as follows. We want the model to select the correct future for any history. This failing, we want the correct future not to be outguessed by a big margin. That is, we want the correct future to be as close to the model's best guess as possible, when they are not identical. This is what D_1 attempts to capture. The second measure D_2 attempts to enforce that correct future clearly stands above all other competing hypotheses even when it is the model's best guess, and that correct future comes close to the best guess when they are not identical. That is, D_2 encourages the model to discriminate all competing hypotheses against the correct hypothesis. All this is in relation to *all* of training data, on average.

This is related to ℓ_∞-version of minimum classification error discriminant described in [1]. In deed, if $\mathcal{F} = \{1, 2, \cdots, M\}$, and if we define $g_i(h) := \log P(i|h)$,

*PLEASE DO NOT DISTRIBUTE. SUBMITTED TO ICASSP-99

then the misclassification measure in Equation (13) of [1] is identical to the summand in (2) above, modulo the sign of the objective function. For further identification of the approaches, we observe that $\ell_k(d_k) := d_k$ leads to $\Leftrightarrow D_2$ and $\ell_k(d_k) := \max(0, d_k)$ leads to $\Leftrightarrow D_1$. However, in [1], the conditional probabilities themselves are not used in discriminant functions. In [2], class conditional likelihood functions are used as discriminant functions; however, where they use $P(h|f)$, we use $\log P(f|h)$. It is important to note that we define our objective function directly in terms of *a posteriori* probabilities. We believe that using *a posteriori* probabilities directly is better [3].

The theory described here applies equally well to both measures of discrimination. So, we focus only on D_1 for the sake of notational simplicity, referring to it as *the* discrimination. We even drop the subscript and write D_1 as D from now on.

The definition of discrimination is applicable to any arbitrary conditional probability distribution. Given a class of models, the goal then would be to choose the model that maximizes discrimination. We consider maximizing discrimination on a class of models that is well-founded in the maximum entropy (minimum Kullback-Leibler distance) framework. This class is a parametric family of exponential models described below. On this class of models, the discrimination turns out to be globally convex in the parameter (in R^n).

3. THE MODEL CLASS

The exponential family of models has three components: a prior distribution, a set of features, and weights associated with these features. The prior distribution models any prior knowledge we may have about the underlying problem. When there is no prior knowledge, the prior distribution is uniform. Typically, the features are binary questions on history and future. Formal description of the family now follows.

We start with a given conditional distribution $P_0(f|h)$, called the *prior* distribution. We are also given a vector function $\phi(h, f)$ whose components are real-valued. We call ϕ the feature function.

We consider a family $\mathcal{P}$ of exponential models parametrized by $\lambda \in R^n$ as below:

$$P_{\lambda,\phi}(f|h) = \frac{P_0(f|h)e^{\lambda\phi(h,f)}}{Z_{\lambda,\phi}(h)},$$

with the normalization factor $Z_{\lambda,\phi}$ given by

$$Z_{\lambda,\phi}(h) := \sum_f P_0(f|h)e^{\lambda\phi(h,f)}.$$

Recall that $\mathcal{P}$ arises in the dual formulation of the minimum Kullback-Leibler "distance" problem (maximum entropy problem if the prior is uniform) and that optimal solution to the primal problem is the maximum likelihood solution in the dual formulation. In other words, one chooses λ to maximize the likelihood of training data according to the model $P_{\lambda,\phi}$. In this paper, we *start with* the dual formulation and replace likelihood of training data with discrimination as the objective function. Also recall that discrimination is an upper bound to likelihood. Finally, even though we sometimes say that $\mathcal{P}$ is "centered on" P_0, the prior P_0 is not a distinguished member of the family; $\mathcal{P}$ can be reparametrized around any other member. We use this reparametrization argument later in a proof.

4. MAIN RESULTS

We now look at the discrimination of these models. We write $P_{\lambda,\phi}$ as P_λ or even as P when convenient. We abuse the notation and write $D(P_{\lambda,\phi})$ as $D(\lambda)$, since with ϕ and P_0 fixed D is a function of only λ.

For any model P, we are interested in the best guess of future. However, it is possible that two different f's can achieve the maximum. A *selector* is the process that assigns for each h *an* $\hat{f}_P(h)$ that maximizes $P(\cdot|h)$. A model can have several selectors associated with it. We denote a particular selector's best guess by

$$\hat{f_P}(h) := \arg\max_{f \in \mathcal{F}} P(f|h).$$

Simplifying the notation, we will write $\hat{f}_{P_{\lambda,\phi}}$ as $\hat{f}_\lambda$. For a fixed h_i, $\hat{f}_\lambda$ can be considered a function of λ. In spite of $\hat{f}_\lambda$ being discontinuous, it turns out that D is a *convex* function of λ. We will also develop an iterative algorithm to find the maximizer of D that involves two steps in each iteration: *selection* and *maximization*. In the selection step, we choose a selector for the current model. In the maximization step, we maximize the objective function over all λ that do not change the selector. The maximization is done by solving a *linear* programming problem.

Notation.

$$\psi_i(\lambda) := \log P_0(\hat{f}_\lambda(h_i)|h_i)e^{\lambda\phi(h_i,\hat{f}_\lambda(h_i))}$$

$$c := \sum_i \log P_0(f_i|h_i)$$

$$d := \sum_i \phi(h_i, f_i)$$

$$\hat{d}(\lambda) := \sum_i \phi(h_i, \hat{f}_\lambda(h_i))$$

The notation for $\hat{d}$ hopefully suggests that the defining sum is a function of the selector as well.

Notice that

$$\hat{f}_\lambda(h_i) = \arg\max_f \frac{P_0(f|h_i)e^{\lambda\phi(h_i,f)}}{Z_{\lambda,\phi}(h_i)}$$

$$= \arg\max_f P_0(f|h_i)e^{\lambda\phi(h_i,f)}$$

which implies that

$$P_0(\hat{f}_\lambda(h_i)|h_i)e^{\lambda\phi(h_i,\hat{f}_\lambda(h_i))} = \max_f P_0(f|h_i)e^{\lambda\phi(h_i,f)}.$$

With this notation, we have

$$\begin{aligned} D(\lambda) &= c + \lambda d \Leftrightarrow \sum_i \psi_i(\lambda) \\ &= c \Leftrightarrow \sum_i \log P_0(\hat{f}_\lambda(h_i)|h_i) + \lambda[d \Leftrightarrow \hat{d}(\lambda)] \\ &= D(0) + \lambda[d \Leftrightarrow \hat{d}(\lambda)] + \sum_i \log \frac{P_0(\hat{f}_0(h_i)|h_i)}{P_0(\hat{f}_\lambda(h_i)|h_i)} \end{aligned}$$

from which follows a useful lower bound for $D(\lambda)$.

Lemma 1. $D(\lambda) \geq D(0) + \lambda[d \Leftrightarrow \hat{d}(\lambda)]$

Proof. Follows from $P_0(\hat{f}_0(h_i)|h_i) \geq P_0(\hat{f}_\lambda(h_i)|h_i)$. □

Proposition 1. $D(\lambda)$ is $\cap$-convex in λ.

Proof. It is enough to show that $\psi_i(\lambda)$ is $\cup$-convex in λ. Fix λ_1, λ_2. For any $\beta \in [0,1]$, $\psi_i(\beta\lambda_1 + (1 \Leftrightarrow \beta)\lambda_2)$

$$\begin{aligned} &= \max_f \quad \log P_0(f|h_i)e^{\beta\lambda_1\phi + (1-\beta)\lambda_2\phi} \\ &= \max_f \quad \beta\lambda_1\phi + (1 \Leftrightarrow \beta)\lambda_2\phi + \log P_0(f|h_i) \\ &= \max_f \quad \beta(\lambda_1\phi + \log P_0(f|h_i)) \; + \\ &\qquad (1 \Leftrightarrow \beta)(\lambda_2\phi + \log P_0(f|h_i) \\ &= \max_f \quad \beta \log P_0(f|h_i)e^{\lambda_1\phi} \; + \\ &\qquad (1 \Leftrightarrow \beta) \log P_0(f|h_i)e^{\lambda_2\phi} \\ &\leq \beta \max_f \quad \log P_0(f|h_i)e^{\lambda_1\phi} \; + \\ &\qquad (1 \Leftrightarrow \beta) \max_f \quad \log P_0(f|h_i)e^{\lambda_2\phi} \\ &= \beta\psi_i(\lambda_1) + (1 \Leftrightarrow \beta)\psi_i(\lambda_2), \end{aligned}$$

which shows that ψ_i is convex. □

In a similar way we can show that the set of λ such that $\hat{f}_\lambda(h_i) = \hat{f}_0(h_i)$ is convex. But it turns out that the set is more than convex; it is a convex polyhedron. We are intereseted in this set because it is the primary ingredient in our optimization algorithm.

Lemma 2. The set $\{\lambda : \hat{f}_\lambda(h_i) = \hat{f}_0(h_i) \; \forall i\}$ is a convex polyhedron.

Proof. We show that the set $\{\lambda : \hat{f}_\lambda(h_i) = \hat{f}_0(h_i)\}$ for a fixed i is a convex polyhedron. The proof is constructive. We have $\hat{f}_\lambda(h_i) = \hat{f}_0(h_i)$

$$\begin{aligned} &\Leftrightarrow \quad P_\lambda(f|h_i) \leq P_\lambda(\hat{f}_0(h_i)|h_i) \quad \forall f \\ &\Leftrightarrow \quad P_0(f|h_i)e^{\lambda\phi(h_i,f)} \leq P_0(\hat{f}_0(h_i)|h_i)e^{\lambda\phi(h_i,\hat{f}_0(h_i))} \\ &\qquad \forall f : P_0(f|h_i) \neq 0 \\ &\Leftrightarrow \quad \lambda[\phi(h_i,f) \Leftrightarrow \phi(h_i,\hat{f}_0(h_i))] \leq \log \frac{P_0(\hat{f}_0(h_i)|h_i)}{P_0(f|h_i)} \\ &\qquad \geq 0 \quad \forall f : P_0(f|h_i) \neq 0. \end{aligned}$$

So, $\{\lambda : \hat{f}_\lambda(h_i) = \hat{f}_0(h_i)\}$ is a polyhedron for each i. □

With binary features, for any f, $\phi(h_i,f) \Leftrightarrow \phi(h_i,\hat{f}_0(h_i)$ takes values from a set of 3^n vectors where n is the size of ϕ. It now follows that

$$\{\lambda : \hat{f}_\lambda(h_i) = \hat{f}_0(h_i) \; \forall i\} = \{\lambda : A\lambda \leq b\}$$

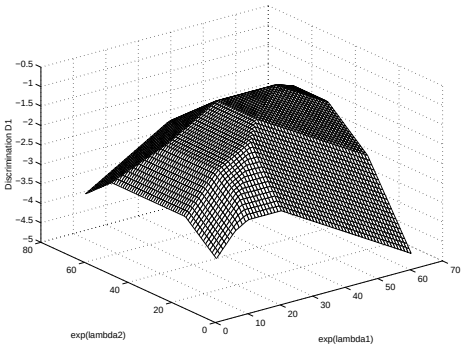
where b is an m-vector ($m \leq 3^n$) with nonnegative components and A is an $m \times n$ matrix. Recall that

$$D(\lambda) = D(0) + \lambda[d \Leftrightarrow \hat{d}(\lambda)] + \sum_i \log \frac{P_0(\hat{f}_0(h_i)|h_i)}{P_0(\hat{f}_\lambda(h_i)|h_i)}.$$

The utility of the set $\Lambda := \{\lambda : \hat{f}_\lambda(h_i) = \hat{f}_0(h_i) \; \forall i\}$ lies in the fact that $D(\lambda)$ is *linear* on this set: First notice that $\hat{d}(\lambda) = \hat{d}(0)$ on Λ. So, $D(\lambda) = D(0) + \lambda[d \Leftrightarrow \hat{d}(0)]$ on Λ. We just made

Observation 1. D is piece-wise linear on R^n.

Here, each piece is a polytope. The objective function for an example problem on R^2 is shown below.



We can think of R^n being partioned into adjacent polytopes. The objective function is a continuous piece-wise linear function on these polytope pieces. Therefore, maximizing $D(\lambda)$ on Λ is reduced to a linear programming problem. We now characterize global maximizers.

Proposition 2. λ_* maximizes $D(\lambda)$ on R^n if and only if there is a selector $\hat{d}(\cdot)$ such that $\hat{d}(\lambda_*) = d$.

Proof. Without loss of generality we may assume that $\lambda_* = 0$, since we can always reparametrize the models around λ_*.

(If) So, we have a selector such that $\hat{d}(0) = d$. On the polytope on which the selector does not change,

we have $D(\lambda) = D(0) + \lambda[d \Leftrightarrow \hat{d}(0)] = D(0)$. That is, $D(\cdot)$ is constant on the polytope, and $D(0)$ is a local maximum. Since D is convex, $D(0)$ is also the global maximum.

(Only if)

Case 1. Suppose $D'(0)$ exists at 0. Then, there is a polytope that contains 0 on which the selector does not change. On this polytope, $D'(\lambda) = d \Leftrightarrow \hat{d}(0) = 0$.

Case 2. (Sketch) $D'(0)$ does not exist at 0; that is, 0 is on the boundary between polytopes. Suppose there is a component k of $(d \Leftrightarrow \hat{d}(\lambda))$ that does not switch signs as we move from one polytope to another. Suppose the sign is +ve. Then, we must be able to increase the objective function by a slight perturbation on the optimal λ i.e. [0 ...0 $+\epsilon$ 0 .. 0] where k-th component is ϵ.

We now state the algorithm to find a maximizer of $D(\lambda)$.

Iter 0. Set $\lambda^{(0)} = 0$.

Iter k+1.

Selection: Choose a selector such that $0 \neq c^{(k)} := d \Leftrightarrow \hat{d}(\lambda^{(k)})$ and at least one selection of maximizer for a history changes from the previous iteration.

Maximization: 1. Compute the polyhedron

$$\Lambda^{(k)} := \{\lambda : \hat{f}_{\lambda}(h_i) = \hat{f}_{\lambda^{(k)}}(h_i) \quad \forall i\}.$$

2. Solve for λ_* where

$$\lambda_* := \arg \max_{\lambda \in \Lambda^{(k)}} c^{(k)} \lambda$$

3. Set $\lambda^{(k+1)} = \lambda^{(k)} + \lambda_*$.

4. Stop if $c^{(k)}\lambda_*$ is less than preset precision. Else continue iteration.

Unlike the maximum entropy (minimum Kullback-Leibler distance) framework, optimal solution is not necessarily unique. We used the discrimination function described here on a natural language understanding task [3]. We obtained best results on the translation task by using a combination of maximum entropy and maximum discrimination and with feature selection. The framework described here is amenable to using the two methods in combination, since we can choose a model developed in one framework as the prior model for the other framework.

ACKNOWLEDGMENT

I thank S. Dharanipragada, R.T. Ward, and S. Roukos for many enjoyable and helpful discussions on this topic.

REFERENCES

[1] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification", *IEEE Trans. Signal Processing*, Vol. 40, No. 12, Dec. 1992, pp. 3043-3054.

[2] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition", *IEEE Trans. Speech and Audio Processing*, vol 5, No. 3, May 1997, pp. 257-265.

[3] K. A. Papineni, R. T. Ward, and S. Roukos, "Maximum likelihood and discriminative training of direct translation models," *Proc. ICASSP-98*, vol I, pp. 189-192, Seattle, 1998.