RC 21815 (98171) 3 August 2000 **ComputerScience**

IBM Research Report

A Generalized Kulback-Leibler Distance and its Minimization

Kishore Papineni
IBM T. J. Watson Research Center
P. O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Haifa - T. J. Watson - Tokyo - Zurich

A generalized Kullback-Leibler distance and its minimization

Kishore Papineni

In this note, we discuss a generalization of the concept of maximum entropy to nonnegative functions that are not necessarily probability distributions. We first define an extension of training data likelihood according to a probability distribution to a generalized likelihood according to a nonnegative function. We then define an exponential family of nonnegative functions that has three components: a fixed prior nonnegative function, a set of features, and their corresponding real weights. We consider a well-posed optimization problem on the family. We show that the objective function is globally convex, and that it corresponds to the dual of a version of Kullback-Leibler distance minimization (or entropy maximization) suitably generalized to nonnegative functions. We then provide a closed-form solution to a one-dimensional optimization problem with a single scalar binary feature function. We propose a version of the improved iterative scaling algorithm to solve a general multi-dimensional optimization problem and prove its convergence to the unique solution.

1 Introduction

We consider modeling training data presented in the form of a set of (history, future) pairs. Each pair is called a training event. In a classification problem, the history in a training event is the observation and the future is the true class. Suppose the set of all futures, $\mathcal{F}$, is finite and is fixed a priori. A traditional model of the data is a conditional probability distribution on the futures conditioned on the history. Such a conditional probability distribution may be realized as a decision tree, or as an exponential distribution constructed in terms of features and their weights in the conditional maximum entropy (minimum Kullback-Leibler "distance") framework. Given a model of the training data, one typically applies it to new unseen histories and rank-orders the futures by assigning nonnegative scores (probabilities) to the futures. Some applications do not require that the model be a true probability distribution, i.e., that the scores add up to one. A classification problem may be one such instance. Another example is a natural language understanding system that selects one formal-language sentence (out of finitely many) as the meaning of a natural language sentence. A note on the terminology: in some disciplines, histories are called feature vectors. A feature means a different thing in this note. Also, a future is simply the predicted variable and does not connote 'the next element' in a sequence unless we are modeling a time series.

When the model is a conditional probability distribution $P(f|h)$, a measure of its goodness is the likelihood of the training data according to the model. Suppose the training data is given as $\{(h_i, f_i)\},\ i = 1, 2, \cdots, T$. The training data log-likelihood according to P is then

$$\sum_{i=1}^{T} \log P(f_i|h_i) \ .$$

We extend these standard notions to nonnegative functions $Q : (h, f) \mapsto [0, \infty)$. The function

Q is written suggestively as $Q(f|h)$. Also, for any function $g : (h, f) \mapsto R^k, k \geq 1$,

$$Q[g] := \sum_{i=1}^{T} \sum_{f \in \mathcal{F}} Q(f|h_i) g(h_i, f).$$

We define a generalized likelihood of training data according to Q as below:

$$G(Q) \ := \ T - Q[1] + \sum_{i=1}^{T} \log Q(f_i|h_i) \ .$$

In case Q is a probability distribution, $G(Q)$ clearly coincides with the standard likelihood of training data according to Q. The generalized likelihood is well-behaved with respect to scaling: For arbitrary $\zeta := [\zeta_1, \zeta_2, \cdots, \zeta_T]$, define $Q_\zeta(f|h_i) := \zeta_i Q(f|h_i), \ \forall i$. Then

$$\max_{\zeta \in R^T} G(Q_\zeta)$$

is finite and is achieved by

$$\zeta_i^* = \frac{1}{\sum_f Q(f|h_i)}.$$

That is, the optimal solution to the scaling problem results in a probability distribution.

Now that we have a well-defined figure of merit of a general nonnegative model of training data, we can consider maximizing the figure over a family of models. To that end, we start with an *a priori* fixed nonnegative function $Q_0(f|h)$, called the *prior model*. We also fix *a priori* a vector feature function $\phi : (h, f) \mapsto R^n$. We note that ϕ is a column vector. We now consider an "exponential" family $\mathcal{Q}$ of models parametrized by a row vector $\lambda \in R^n$ as below.

$$\mathcal{Q} \ := \ \{Q(f|h) := Q_0(f|h) e^{\lambda \phi(h,f)}, \ \forall h, f\}$$

For a *scalar* $g(h, f)$, let

$$(g \circ Q)(f|h) := Q(f|h) e^{g(h,f)}, \ \forall h, f.$$

Using this notation, we have

$$\mathcal{Q} \ := \ \{\lambda \phi \circ Q_0\}.$$

$\mathcal{Q}$ is easily seen to be convex and is analogous to the exponential family of probability distributions that arises in the maximum entropy (minimum Kullback-Leibler framework) framework. We simply relaxed the requirement on normalization. Clearly, once $\mathcal{Q}$ is fixed, Q_0 is no longer a distinguished member of $\mathcal{Q}$ in that $\mathcal{Q}$ can be reparametrized around any other member.

We now have a class of models and an objective function on them. We consider optimizing the objective function on the class of models. We first denote the empirical feature counts by d as below.

$$d \ := \ \sum_{i=1}^{T} \phi(h_i, f_i)$$

It is then routine to show that for all $Q \in \mathcal{Q}$

$$G(Q) = G(Q_0) + Q_0[1] - T + T + \lambda d - \sum_i \sum_f Q_0(f|h_i) e^{\lambda \phi(h_i, f)} \ .$$

Defining

$$L(\lambda) := T + \lambda d - \sum_i \sum_f Q_0(f|h_i) e^{\lambda \phi(h_i, f)},$$

we see that

$$\sup_{Q \in \mathcal{Q}} G(Q)$$

is equivalent to

$$\sup_{\lambda \in R^n} L(\lambda) \qquad (OPT1).$$

Observation 1. $L(\lambda)$ is globally strictly convex in λ.

Observation 2. The following is necessary and sufficient for a $\lambda_* \in R^n$ to be optimal:

$$d = \sum_i \sum_f Q_0(f|h_i) e^{\lambda_* \phi(h_i, f)} \phi(h_i, f)$$

Remark. In the conditional maximum entropy framework, one starts with an exponential family of *probability distributions* and maximizes the training data likelihood according to these distributions. This formulation results in maximizing the following function over λ,

$$\lambda d - \sum_i \log \sum_f Q_0(f|h_i) e^{\lambda \phi(h_i, f)},$$

which is solved by improved iterative scaling [1].

In the next section, we show that the maximization problem (OPT1) is indeed the dual of a generalized Kullback-Leibler distance minimization problem. In section 3, we solve the single-parameter optimization problem explicitly when ϕ is a scalar binary feature. We then compare the scalar solution to that in the maximum entropy framework. In Section 4, we propose an iterative algorithm for the multiparameter problem with vector features and show that it converges to the unique optimal solution.

2 A generalization of Kullback-Leibler distance

Given the training data as in Section 1, the following modification of Kullback-Leibler "distance" between two conditional *probability distributions* is standard in statistical natural language processing. Assume $P_1(f|h) = 0$ whenever $P_2(f|h) = 0$.

$$D_{KL}(P_1, P_2) \ := \ \sum_{i=1}^{T} \sum_{f \in \mathcal{F}} P_1(f|h_i) \log \frac{P_1(f|h_i)}{P_2(f|h_i)} \ .$$

We consider the following generalization to *nonnegative* functions, $Q_1, Q_2 \geq 0$.

$$D_G(Q_1, Q_2) \ := \ T - Q_1[1] + \sum_{i=1}^{T} \sum_{f \in \mathcal{F}} Q_1(f|h_i) \log \frac{Q_1(f|h_i)}{Q_2(f|h_i)}.$$

Clearly, if P_1 is a probability distribution, then $D_G(P_1, Q_2) = D_{KL}(P_1, Q_2)$.

We now discuss a number of properties of D_G. The next lemma is routine.

Lemma 1. For any $Q, Q_0 \geq 0$, we have

(i) $D_G(Q, Q) = T - Q[1]$

(ii) $D_G(Q, Q_0) - D_G(Q, g \circ Q_0) = Q[g]$

Lemma 2. $D_G(Q_1, Q_2)$ is strictly convex in Q_1 and Q_2 separately.

Proof. Use log-sum inequality (p 27, [2]).

Recall that

$$d \ := \ \sum_{i=1}^{T} \phi(h_i, f_i).$$

A priori, fix a $Q_0(f|h) \geq 0$, called the prior model. Define a family $\mathcal{R}$ as below:

$$\mathcal{R} \ := \ \{Q \geq 0 : Q[\phi] = d\}$$

We now pose a (primal) optimization problem.

$$\inf_{Q \in \mathcal{R}} D_G(Q, Q_0)$$

Formulating the Lagrangian $L(Q, \lambda)$ as usual, differentiating it with respect to $Q(\cdot|\cdot)$ and setting to 0, we conclude that $Q(f|h)$ is necessarily of the form:

$$Q(f|h) = Q_0(f|h) e^{\lambda \phi(h,f)}.$$

Substituting this structural form into $L(Q, \lambda)$, we get a function of λ alone:

$$L(\lambda) = T + \lambda d - \sum_i \sum_f Q_0(f|h_i) e^{\lambda \phi(h_i, f)}.$$

Therefore, the primal problem is equivalent to the following dual:

$$\sup_{\lambda \in R^n} L(\lambda)$$

which was shown in Section 1 to be equivalent to

$$\sup_{Q \in \mathcal{Q}} G(Q).$$

Finally, the distinction between $\mathcal{Q}$ and its closure is not important and we will not make such a distinction in what follows.

3 The Scalar Binary Feature Problem

The scalar feature problem is important since we select good features from a pool of possibly hundreds of thousands of features. Feature selection involves solving a scalar feature problem for every feature in the pool.

Recall that the optimal λ_* satisfies

$$d = \sum_i \sum_f Q_0(f|h_i) e^{\lambda_* \phi(h_i, f)} \phi(h_i, f).$$

If the feature is binary, we have

$$\phi(h,f)e^{\lambda\phi(h,f)} = \phi(h,f)e^{\lambda}$$

from which follows

$$d = e^{\lambda_*}\sum_{i=1}^{T}\sum_{f\in\mathcal{F}} Q_0(f|h_i)\phi(h_i,f).$$

Defining

$$d_0 = \sum_{i=1}^{T}\sum_{f\in\mathcal{F}} Q_0(f|h_i)\phi(h_i,f),$$

we have

$$\lambda_* = \log\frac{d}{d_0}.$$

The optimal *gain* $L(\lambda_*) - L(0)$ is easily shown to be

$$L(\lambda_*) = d_0 - d + d\log\frac{d}{d_0}.$$

4 Non-overlapping binary features

We say that a set of binary features $\{\phi_k(h,f)\}$ is non-overlapping if $\sum_k \phi_k(h,f)$ is also binary-valued. Consider the feature vector $\phi_1^n := [\phi_1\ \phi_2 \cdots \phi_n]'$ whose components are binary and non-overlapping. Suppose Q_{n-1} solves OPT_{n-1} below:

$$\min_{Q\geq 0} D_G(Q,Q_0)$$

subject to

$$Q[\phi_1^{n-1}] = d_1^{n-1} := \sum_{i=1}^{T}\phi_1^{n-1}(h_i,f_i).$$

Likewise, suppose Q_n solves OPT_n. Then for some $\Delta := [\delta_1\ \delta_2\ \cdots \delta_{n-1}] \in R^{n-1}$ and for some $\lambda_n \in R$ we have

$$Q_n = \lambda_n\phi_n\ \circ\ \Delta\phi_1^{n-1}\ \circ\ Q_{n-1}.$$

Since Q_n solves OPT_n, the following holds in particular for $k = 1, 2, \cdots, n-1$:

$$d_k = \sum_{i=1}^{T}\sum_{f} Q_n(f|h_i)\phi_k(h_i,f) = \sum_{i=1}^{T}\sum_{f} Q_{n-1}(f|h_i)e^{\lambda_n\phi_n(h_i,f)}e^{\Delta\phi_1^{n-1}(h_i,f)}\phi_k(h_i,f).$$

But $\phi_k(h_i,f) = 1$ implies that $\Delta\phi_1^{n-1}(h_i,f) = \delta_k$ and that $\phi_n(h_i,f) = 0$. From this, it follows that

$$d_k = \sum_{i=1}^{T}\sum_{f} Q_{n-1}(f|h_i)e^{\delta_k}\phi_k(h_i,f) = e^{\delta_k}\sum_{i=1}^{T}\sum_{f} Q_{n-1}(f|h_i)\phi_k(h_i,f).$$

Since Q_{n-1} solves OPT_{n-1}, we have

$$d_k = e^{\delta_k}d_k,$$

from which we conclude $\delta_k = 0$ for $k = 1, 2, \cdots, n-1$. This means that

$$Q_n = \lambda_n\phi_n\ \circ\ Q_{n-1}.$$

We therefore obtain λ_n by solving the scalar binary feature problem whose solution is explicit:

$$\lambda_n = \log \frac{d_n}{Q_{n-1}[\phi_n]}.$$

However, whenever $\phi_n(h,f) = 1$, we have $Q_{n-1}(h,f) = Q_0(h,f)$ since the features are non-overlapping. This means that

$$\lambda_n = \log \frac{d_n}{Q_0[\phi_n]}.$$

Therefore, we have the following lemma.

Lemma 4. The multiparameter optimal solution to n non-overlapping binary features is simply obtained by solving n scalar binary feature problems independently, *all with the same prior* Q_0. That is, Opt $[\lambda_k]_1^n = [$ Opt $\lambda_k]_1^n$.

The following lemma is also straightforward.

Lemma 5. For non-overlapping binary features, the multi-parameter optimal gain is the sum of the single parameter optimal gains.

5 The General Multiparameter Problem

In this section, we propose a version of the iterative scaling algorithm of [1] for optimizing G(Q) over $\mathcal{Q}$ and prove that the sequence of models proposed by the algorithm converges to the (unique) optimal solution.

Recall the definition of $g \circ Q$ for any scalar function $g(h,f)$ and for any $Q \geq 0$:

$$(g \circ Q)(f|h) := Q(f|h)e^{g(h,f)}.$$

Since ϕ is fixed, we abuse the notation and write $(\lambda\phi) \circ Q$ as simply $\lambda \circ Q$. Also recall the definition of $G(Q)$:

$$G(Q) := T - Q[1] + \sum_{i=1}^{T} \log Q(f_i|h_i).$$

Let $\Delta := \{Q \geq 0\}$.

Definition. A function $A : R^n \times \Delta \mapsto R$ is an auxiliary function for G if

(1) For all $Q \in \Delta$ and $\lambda \in R^n$

$$G(\lambda \circ Q) \geq G(Q) + A(\lambda, Q)$$

(2) $A(\lambda, Q)$ is continuous in $Q \in \Delta$ and C^1 in λ with $A(0,Q) = 0$ and

$$\frac{d}{dt}|_{t=0} A(t\lambda, Q) = \frac{d}{dt}|_{t=0} G((t\lambda) \circ Q).$$

We use an auxiliary function A to construct an iterative algorithm for maximizing G. We start with $Q^{(0)} := Q_0$ and recursively define $Q^{(k+1)}$ by

$$Q^{(k+1)} = \lambda^{(k)} \circ Q^{(k)}$$

with

$$\lambda^{(k)} := \arg\max_{\lambda} A(\lambda, Q^{(k)}).$$

Proposition 1. Suppose $Q^{(k)}$ is any sequence in Δ with

$$Q^{(0)} = Q_0 \quad \text{and} \quad Q^{(k+1)} = \lambda^{(k)} \circ Q^{(k)}$$

where $\lambda^{(k)} \in R^n$ satisfies

$$A(\lambda^{(k)}, Q^{(k)}) = \sup_{\lambda} A(\lambda, Q^{(k)}).$$

Then $G(Q^{(k)})$ increases monotonically to $\max_{Q \in \mathcal{Q}} G(Q)$ and $Q^{(k)}$ converges to the unique $Q_* := \arg\max_{Q \in \mathcal{Q}} G(Q)$.

Proof. Suppose m is a cluster point of $Q^{(k)}$. Then, there is a subsequence $Q^{(k_l)}$ that converges to m. Then, for any λ,

$$A(\lambda, Q^{(k_l)}) \leq A(\lambda^{(k_l)}, Q^{(k_l)}) \leq G(Q^{(k_l+1)}) - G(Q^{(k_l)}) \leq G(Q^{(k_{l+1})}) - G(Q^{(k_l)}).$$

Taking limits, we see that $A(\lambda, m) \leq 0$ for any λ. Since $A(0, m) = 0$, this implies that $\lambda = 0$ is a maximum of $A(\lambda, m)$ so that

$$0 = \frac{d}{dt}|_{t=0} A(t\lambda, m) = \frac{d}{dt}|_{t=0} G((t\lambda) \circ m), \quad \forall \lambda \in R^n.$$

This implies that $m \in \mathcal{R}$ which in turn implies that m optimizes $G(Q)$. Since the objective function is strictly convex, there is at most one optimizer. Therefore, $m = Q_*$. This in turn implies that $Q^{(k)}$ itself converges to Q_*. □

Finally, defining

$$\phi_\#(h, f) := \sum_{j=1}^{n} \phi_j(h, f),$$

we note that

$$A(\lambda, Q) := Q[1] + \lambda d - Q[\sum_{j=1}^{n} \frac{\phi_j(h, f)}{\phi_\#(h, f)} e^{\lambda_j \phi_\#(h,f)}]$$

is an auxiliary function for G in this note.

References

[1] S. Della Pietra, V. Della Pietra, and J. Lafferty, "Inducing features of random fields," IEEE Transactions on Pattern Analysis and Machine Intelligence, 19(4), April 1997, pp. 380-393.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory,* John Wiley & Sons, New York, 1991.