Growth Transform for Conditional Maximum Likelihood Estimation of Log-Linear Models

Milind Mahajan and Ciprian Chelba
{milindm,chelba}@microsoft.com

June 2002

Technical Report
MSR-TR-2002-65

We present a method for conditional maximum likelihood estimation of log-linear models that employs a well known technique relying on a generalization of the Baum-Eagon inequality from polynomials to rational functions.

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

1 Introduction

In many practical applications one seeks to model a conditional probability $P(y|x), y \in \mathcal{Y}, x \in \mathcal{X}$. A common situation is that in which we identify a set of features deemed relevant for building the model; the features are non-negative functions — usually indicator functions $f(x,y) : \mathcal{X} \times \mathcal{Y} \rightarrow \{0,1\}$. Let $\mathcal{F} = \{f_k, k = 1 \ldots F\}$ be the set of features chosen for building a particular model $P(y|x)$.

The conditional log-linear model one wishes to estimate is of the form:

$$
\begin{aligned}
P(y|x) &= Z(x)^{-1} \prod_{i=1}^{F} {\lambda_i}^{f_i(x,y)} \qquad (1) \\
Z(x)^{-1} &= \sum_{z \in \mathcal{Y}} \prod_{i=1}^{F} {\lambda_i}^{f_i(x,z)}
\end{aligned}
$$

2 Conditional Maximum Likelihood Estimation of Log-Linear Models

As noted in [1] and [2], one can assume that $\sum_i f_i(x,y) = M, \forall (x,y) \in \mathcal{X} \times \mathcal{Y}$ without restricting the generality of the family of log-linear models under consideration.

Under these conditions the feature weights λ_i can be normalized such that they become a probability distribution over indices $i = 1 \ldots F$:

- $\lambda_i >= 0, \forall i \in 1 \ldots F$
- $\sum_{i=1}^{F} \lambda_i = 1$

It is desirable to estimate the feature weights (probabilities) λ_i such that the conditional likelihood $H(\mathcal{T}; \underline{\lambda}) = \prod_{j=1}^{T} P(y_j|x_j)$ assigned by the model to a set of training samples $\mathcal{T} = \{(x_1, y_1) \ldots (x_T, y_T)\}$ is maximized:

$$\underline{\lambda}^* = \arg\max_{\underline{\lambda}} H(\mathcal{T}; \underline{\lambda}) \qquad (2)$$

It is easy to note that $H(\mathcal{T}; \underline{\lambda})$ is a ratio of two polynomials with real coefficients, each defined over a domain D of probability distributions $\underline{\lambda}$ over indices $i = 1 \ldots F$:

$$D = \{\underline{\lambda} : \lambda_i >= 0, \forall i \in 1 \ldots F, \sum_{i=1}^{F} \lambda_i = 1\}$$

Following the development in [3] one can iteratively estimate the feature weights using a growth transform for rational functions on the domain D. The reestimation equations take the form:

$$\widehat{\lambda_i} = N^{-1} \lambda_i \left(\frac{\partial \log H(\mathcal{T}; \underline{\lambda})}{\partial \lambda_i} + C_{\underline{\lambda}}\right) \qquad (3)$$

$$N = \sum_{i=1}^{F} \lambda_i (\frac{\partial \log H(\mathcal{T}; \underline{\lambda})}{\partial \lambda_i} + C_{\underline{\lambda}})$$

where $C_{\underline{\lambda}}$ is chosen such that

$$\frac{\partial \log H(\mathcal{T}; \underline{\lambda})}{\partial \lambda_i} + C_{\underline{\lambda}} > 0, \forall i$$

Calculating the partial derivatives in Eq.(3) we obtain:

$$\widehat{\lambda_i} = N^{-1} \left\{ \lambda_i \cdot C_{\underline{\lambda}} + T \left(E_{f(X)f(Y|X)}[f_i(X,Y)] - E_{f(X)p(Y|X;\underline{\lambda})}[f_i(X,Y)] \right) \right\} \quad (4)$$

$$N = \sum_{i=1}^{F} \lambda_i \cdot C_{\underline{\lambda}} + T \left(E_{f(X)f(Y|X)}[f_i(X,Y)] - E_{f(X)p(Y|X;\underline{\lambda})}[f_i(X,Y)] \right)$$

with $C_{\underline{\lambda}}$ chosen at each iteration such that

$$\frac{T}{\lambda_i} \left(E_{f(X)f(Y|X)}[f_i(X,Y)] - E_{f(X)p(Y|X;\underline{\lambda})}[f_i(X,Y)] \right) + C_{\underline{\lambda}} > 0, \forall i$$

2.1 Comments on Convergence

The fixed points of the growth transform reestimation procedure are the same as the stationary points of the Lagrangian resulting from the maximization of the log-likelihood $L(\underline{\lambda}, \alpha) = \log H(\mathcal{T}; \underline{\lambda}) + \alpha(\sum_{i=1}^{F} \lambda_i - 1)$.

Indeed, at a fixed point of the growth transform we have $\widehat{\lambda_i} = \lambda_i, \forall i = 1 \ldots F$ which is equivalent to

$$\frac{\partial \log H(\mathcal{T}; \underline{\lambda})}{\partial \lambda_i} = const, \forall i = 1 \ldots F \quad (5)$$

Also, the stationary points of the Lagrangian are at:

$$\frac{\partial L(\underline{\lambda}, \alpha)}{\partial \lambda_i} = 0, \forall i = 1 \ldots F$$

which is equivalent to:

$$\frac{\partial \log H(\mathcal{T}; \underline{\lambda})}{\partial \lambda_i} = -\alpha, \forall i = 1 \ldots F \quad (6)$$

It can be easily seen that conditions (5) and (6) are equivalent and thus the fixed points of the growth transform reestimation procedure are the same as the stationary points of the log-likelihood.

Moreover, at fixed points of the growth transform reestimation procedure, the value of *const* in Eq. (5) is 0.

By summing:

$$\frac{\partial \log H(\mathcal{T}; \underline{\lambda})}{\partial \lambda_i} = \frac{T}{\lambda_i} \left(E_{f(X)f(Y|X)}[f_i(X,Y)] - E_{f(X)p(Y|X;\underline{\lambda})}[f_i(X,Y)] \right), \forall i = 1 \ldots F$$

over all indices i and using the linearity of the expectation operator one obtains:

$$const \cdot \sum_{i=1}^{F} \lambda_i = T\left(E_{f(X)f(Y|X)}[\sum_{i=1}^{F} f_i(X,Y)] - E_{f(X)p(Y|X;\underline{\lambda})}[\sum_{i=1}^{F} f_i(X,Y)]\right)$$

which becomes

$$const = T \cdot \left(E_{f(X)f(Y|X)}[M] - E_{f(X)p(Y|X;\underline{\lambda})}[M]\right) = 0$$

after using $\sum_i f_i(x,y) = M, \forall(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $\sum_{i=1}^{F} \lambda_i = 1$.

It follows that the fixed points of the growth transform reestimation procedure are in the intersection of the familiy of log-linear probability distributions

$$P = \{f(X)p(Y|X;\underline{\lambda}) : \lambda_i >= 0, \forall i \in 1 \ldots F, \sum_{i=1}^{F} \lambda_i = 1\} \tag{7}$$

(where $p(Y|X;\underline{\lambda})$ is given by Eqn. (1)) with the linear family

$$Q = \{q(X,Y) : E_{f(X)f(Y|X)}[f_i(X,Y)] = E_{q(X,Y)}[f_i(X,Y)], \forall i \in 1 \ldots F\} \tag{8}$$

It is a well-known fact (see [2]) that the intersection between the two families is unique and represents the maximum-likelihood estimate for probability distributions in P.

We have thus shown that the growth transform reestimation procedure employing the reestimation Eqns. (4) has as a unique fixed point the maximum likelihood distribution in the class of log-linear models P.

3 Conclusions

We have presented an alternative to the familiar Generalized Iterative Scaling [1] algorithm used for conditional maximum likelihood estimation of log-linear models that uses a growth transform for rational functions as derived in [3].

With minor modifications, the reestimation procedure presented can be applied for the estimation of models

$$\begin{aligned} P(y|x) &= Z(x)^{-1} Q(y|x) \prod_{i=1}^{F} {\lambda_i}^{f_i(x,y)} \qquad (9) \\ Z(x)^{-1} &= \sum_{z \in \mathcal{Y}} Q(z|x) \prod_{i=1}^{F} {\lambda_i}^{f_i(x,z)} \end{aligned}$$

that arise in minimum I-divergence estimation.

References

[1] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1470–1480, 1972.

[2] Imre Csizsar, "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," *The Annals of Statistics*, vol. 17, no. 3, pp. 1409–1413, 1989.

[3] P. S. Gopalakrishnan, "An inequality for rational functions with applications to some statistical estimation problems," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 107–113, January 1991.