Conditional Maximum Likelihood Estimation Using Rational Function Growth Transform

Ciprian Chelba*, Alex Acero
Microsoft Research
One Microsoft Way
Redmond, WA 98052
{chelba,alexac}@microsoft.com

We present a study on conditional maximum likelihood (CML) estimation of probability models by means of a well known technique that generalizes the Baum-Eagon inequality [1] from polynomials to rational functions. The main advantage of the rational function growth transform (RFGT) method [5] is that it keeps the model parameter values — probabilities — properly normalized at each iteration. As a case study we apply the technique to discriminatively train a Naïve Bayes (NB) classifier; the same procedure is at the basis of discriminative training of HMMs in speech recognition [6].

The NB model trained under the maximum likelihood (ML) and CML criteria, respectively, is used on a text classification problem. Smoothing is found to be a key component in increasing the classification accuracy. A simple modification of the algorithm increases the convergence speed significantly — as measured by likelihood and classification accuracy increase per iteration — over a straightforward implementation of RFGT. The model trained under the CML criterion achieves a relative improvement of 40% in classification accuracy over its ML counterpart [3].

The NB model can also be re-parameterized as a standard conditional exponential model encoutered in maximum entropy (MaxEnt) estimation [2]. Although the two parameterizations are in principle equivalent and should lead to the same model when trained under CML, the conditional exponential model estimated using improved iterative scaling (IIS) and smoothing with a Gaussian prior [4] outperforms the smoothed NB model estimated using RFGT when evaluated in terms of classification accuracy.

Conditional Naïve Bayes Models In many practical applications one seeks to model a conditional probability $P(y|x), y \in \mathcal{Y}, x \in \mathcal{X}$. We will restrict our attention to using features that are binary valued indicator functions $f(x) : \mathcal{X} \rightarrow \{0, 1\}$, $\overline{f_k(x)} = 1 - f(x)$. Let $\mathcal{F} = \{f_k, k = 1 \ldots F\}$ be the set of features chosen for building a particular model $P(y|x)$. Assuming a NB model for the feature vector and the predicted variable $(\underline{f}(x), y)$ the conditional probability $P(y|x)$ can be calculated as:

$$P(y|x;\underline{\theta}) \quad = \quad Z(x;\underline{\theta})^{-1} \cdot \theta_y \prod_{k=1}^{F} {\theta_{ky}}^{f_k(x)} \overline{\theta}_{ky}^{\overline{f_k(x)}} \tag{1}$$

where: $\theta_y \geq 0, \forall y \in \mathcal{Y}, \sum_{y \in \mathcal{Y}} \theta_y = 1$; $\theta_{ky} \geq 0, \overline{\theta}_{ky} \geq 0, \theta_{ky} + \overline{\theta}_{ky} = 1, \forall k = 1 \ldots F, y \in \mathcal{Y}$ and $Z(x;\underline{\theta})^{-1} = \sum_y P(\underline{f}(x), y)$ is a normalization term.

Rational Function Growth Transform for CML Estimation of Naïve Bayes Models It is desirable to estimate the model parameters $\underline{\theta} = \{\theta_y, \theta_{ky}, \overline{\theta_{ky}}, \forall y \; and \; k\}$ such that the conditional likelihood $H(\mathcal{T};\underline{\theta}) = \prod_{j=1}^{T} P(y_j|x_j)$ assigned by the model to a set of training samples $\mathcal{T} = \{(x_1, y_1) \ldots (x_T, y_T)\}$ is maximized: $\underline{\theta}^* = \arg\max_{\underline{\theta}} H(\mathcal{T};\underline{\theta})$. It is easy to note that $H(\mathcal{T};\underline{\theta})$ is a ratio of two polynomials with real coefficients, each defined over a set $\times$ of probability distributions:

$$\times = \{\underline{\theta} : \theta_y \geq 0, \forall y \in \mathcal{Y} \, and \sum_y \theta_y = 1; \theta_{ky} \geq 0, \overline{\theta}_{ky} \geq 0 \, and \, \theta_{ky} + \overline{\theta}_{ky} = 1, \forall y \in \mathcal{Y}, \forall k = 1 \ldots F\}$$

*Attending Author
Category: Estimation, Prediction, and Sequence Modeling
Preference: 1. Oral 2. Poster

Following the development in [5] one can iteratively estimate the model parameters using a growth transform for rational functions on the domain $\times$. The re-estimation equations take the form:

$$\begin{aligned}
\widehat{\theta_y} &= N^{-1}\theta_y\left(\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \theta_y} + C_{\underline{\theta}}\right) \qquad (2)\\
N &= C_{\underline{\theta}} + \sum_y \theta_y \frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \theta_y}\\
\widehat{\theta_{ky}} &= N_y^{-1}\theta_{ky}\left(\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \theta_{ky}} + C_{\underline{\theta}}\right) \qquad (3)\\
\widehat{\overline{\theta}_{ky}} &= N_y^{-1}\overline{\theta}_{ky}\left(\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \overline{\theta}_{ky}} + C_{\underline{\theta}}\right)\\
N_y &= C_{\underline{\theta}} + \theta_{ky}\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \theta_{ky}} + \overline{\theta}_{ky}\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \overline{\theta}_{ky}}
\end{aligned}$$

where $C_{\underline{\theta}} > 0$ is chosen such that, with $\epsilon > 0$ suitably chosen (see [5] and [3] for details) we have:

$$\begin{aligned}
\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \theta_y} + C_{\underline{\theta}} &> \epsilon, \forall y \qquad (4)\\
\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \theta_{ky}} + C_{\underline{\theta}} &> \epsilon, \forall k \; and \; y\\
\frac{\partial \log H(\mathcal{T};\underline{\theta})}{\partial \overline{\theta}_{ky}} + C_{\underline{\theta}} &> \epsilon, \forall k \; and \; y
\end{aligned}$$

Equivalence with Exponential Models Setting: $f_k(x,y) = f_k(x)\cdot\delta(y)$; $\lambda_{ky} = \log(\frac{\theta_{ky}}{\overline{\theta}_{ky}})$; $\lambda_{0y} = log(\theta_y \cdot \prod_{k=1}^{F}\overline{\theta}_{ky})$ and $f_0(x,y) = f_0(y)$ in Eqn. (1) we obtain the familiar log-linear model arrived at in maximum entropy probability estimation [2]:

$$P(y|x;\underline{\lambda}) = Z(x;\underline{\lambda})^{-1}\cdot exp\left(\sum_{k=0}^{F}\lambda_{ky}f_k(x,y)\right) \qquad (5)$$

where λ_{ky} are free real-valued parameters and $Z(x;\underline{\lambda})$ a normalization term.

References

[1] L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. In *Inequalities*, volume 3, pages 1–8. 1972.

[2] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–72, March 1996.

[3] Ciprian Chelba and Alex Acero. Conditional maximum likelihood estimation of Naïve Bayes probability models. Technical Report to appear, Microsoft Research, Redmond, WA, 2004.

[4] Stanley F. Chen and Ronald Rosenfeld. A survey of smoothing techniques for maximum entropy models. *IEEE Transactions on Speech and Audio Processing*, 8(1):37–50, 2000.

[5] P. S. Gopalakrishnan, Dimitri Kanevski, Arthur Nadas, and David Nahamoo. An inequality for rational functions with applications to some statistical estimation problems. *IEEE Transactions on Information Theory*, 37(1):107–113, January 1991.

[6] Y. Normandin. *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem*. PhD thesis, McGill University, Montreal, 1991.