Pattern Recognition Letters 30 (2009) 1295–1300

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Using continuous features in the maximum entropy model ☆

Dong Yu *, Li Deng, Alex Acero

Microsoft Research, One Microsoft Way, Redmond, WA 98052, United States

ARTICLE INFO

Article history:
Received 16 October 2008
Received in revised form 11 May 2009
Available online 24 June 2009

Communicated by R.C. Guido

Keywords:
Maximum entropy principle
Spline interpolation
Continuous feature
Maximum entropy model
Moment constraint
Distribution constraint

ABSTRACT

We investigate the problem of using continuous features in the maximum entropy (MaxEnt) model. We explain why the MaxEnt model with the moment constraint (MaxEnt-MC) works well with binary features but not with the continuous features. We describe how to enhance constraints on the continuous features and show that the weights associated with the continuous features should be continuous functions instead of single values. We propose a spline-based solution to the MaxEnt model with non-linear continuous weighting functions and illustrate that the optimization problem can be converted into a standard log-linear model at a higher-dimensional space. The empirical results on two classification tasks that contain continuous features are reported. The results confirm our insight and show that our proposed solution consistently outperforms the MaxEnt-MC model and the bucketing approach with significant margins.

1. Introduction

The maximum entropy (MaxEnt) model with moment constraints (MaxEnt-MC) on binary features has been shown effective in natural language processing (NLP) (e.g., Berger et al., 1996), speaker identification (e.g., Ma et al., 2007), statistical language modeling (e.g., Rosenfeld, 1996), text filtering and cleaning (e.g., Yu et al., 2005a), machine translation (e.g., Och and Ney, 2002), phonotactic learning (e.g., Hayes, 2008), visual object classification (e.g., Gong et al., 2004), economic modeling (e.g., Arndt et al., 2002), and network anomaly detection (e.g., Gu et al., 2005). However, it is not very successful when non-binary (e.g., continuous) features are used. To improve the performance, quantization techniques such as bucketing (or binning) have been proposed to convert the continuous features into binary features. Unfortunately, quantization techniques provide only limited performance improvement due to its intrinsic limitations. A coarse quantization may introduce large quantization errors and wash out the gain obtained from using the converted binary features, and a fine quantization may increase the number of model parameters dramatically and introduce parameter estimation uncertainties.

In this paper, we examine the MaxEnt model and the principle behind it. We bring the insight that the key to the success of using the MaxEnt model is providing appropriate constraints. We show that moment constraints on binary features are very strong and fully regularize the distribution of the features. However, moment constraints on continuous features are rather weak and as a result much information contained in the training set is not used by the MaxEnt model. Therefore, using continuous features is less effective than using binary features in the MaxEnt-MC model.

We further discuss how stronger constraints can be included for continuous features by using quantization techniques. We extend the quantization technique to its extreme to introduce the distribution constraint and show that the weights associated with continuous features in the MaxEnt model should not be single values but continuous functions. In other words, the optimization problem is no longer a log-linear problem but a non-linear problem with continuous weighting functions as parameters. We solve this non-linear optimization problem by approximating the continuous weighting function with spline interpolations we recently developed in our variable parameter hidden Markov model (VPHMM) work (Yu et al., 2008, in press). We demonstrate that by using the spline interpolation the optimization problem with non-linear continuous weighting functions can be converted into a standard log-linear problem at a higher-dimensional space where each continuous feature in the original space is mapped into several features. With this transformation, the existing training and testing algorithms (Nocedal, 1980; Riedmiller and Braun, 1993; Malouf, 2002) as well as the recently developed regularization techniques (Chen and Rosenfeld, 1999, 2000; Goodman, 2004; Kazama,

☆ A small portion of this work has been presented at the NIPS 2008 workshop on speech and language: Learning-based methods and systems at Whistler, BC, Canada in December 2008.

* Corresponding author. Fax: +1 425 706 7329.
E-mail addresses: dongyu@microsoft.com (D. Yu), deng@microsoft.com (L. Deng), alexac@microsoft.com (A. Acero).

0167-8655/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.06.005

2004; Kazama and Tsujii, 2005) for the MaxEnt-MC model can be directly applied in this higher-dimensional space making our approach very attractive. We validate our insight and the effectiveness of our approach on two classification tasks that contain continuous features and show that our proposed solution consistently outperforms the MaxEnt-MC model and the quantization-based approach with significant margins.

The rest of the paper is organized as follows. In Section 2, we examine the MaxEnt model and discuss why the MaxEnt model with moment constraints performs well for binary features but not for continuous features. In Section 3, we illustrate that continuous weighting functions (instead of single weight values) should be used for continuous features and propose a solution to the optimization problem that contains continuous weighting functions by approximating the weighting functions with spline interpolations. We validate our insight and demonstrate the new approach's superiority over the MaxEnt-MC and quantization-based approaches empirically on two classification tasks in Section 4, and conclude the paper with discussions on many potential applications in Section 5.

2. The MaxEnt model and constraints

In this section, we examine the MaxEnt principle and the MaxEnt model and explain why the MaxEnt model with moment constraints works well for the binary features but not for the continuous features by showing that the moment constraints on binary features are strong while on continuous features weak.

2.1. The MaxEnt principle and MaxEnt model with moment constraints

We consider a random process that produces an output value y from a finite set Y for an input value x. We assume that a training set $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$ with N samples is given. The training set can be represented with the empirical probability distribution

$$\tilde{p}(x,y)=\frac{\text{number of times that } (x,y) \text{ occur}}{N}. \tag{1}$$

Our goal is to construct a stochastic model that can accurately represent the random process that generated the training set $\tilde{p}(x,y)$. We denote $p(y\mid x)$ as the probability of outputting by y the model when x is given and assume that a set of constraints C is known either from the training data and/or from *a priori* knowledge.

The MaxEnt principle (Guiasu and Shenitzer, 1985) dictates that from all the probability distributions $p(y\mid x)$ that accord with the constraints C, we should select the distribution that is most uniform. Mathematically, we should select the distribution that maximizes the entropy

$$H(p)=-\sum_{x,y}\tilde{p}(x)p(y\mid x)\log p(y|x), \tag{2}$$

over the conditional probability $p(y\mid x)$.

A typical type of constraints used in the MaxEnt model is moment constraints. Assume that a set of M features $f_i(x,y)$, $i=1,\ldots,M$ is available, the moment constraint requires that the moment of the features as predicted from the model should be the same as that observed from the training set. In most cases only the constraints on the first-order moment is used, i.e.,

$$E_p[f_i]=E_{\tilde{p}}[f_i],\quad i=1,\ldots,M, \tag{3}$$

where E_p is the expected value over the distribution p defined as

$$E_p[f_i]=\sum_{x,y}\tilde{p}(x)p(y|x)f_i(x,y), \tag{4}$$

and $E_{\tilde{p}}$ is the expected value over the distribution $\tilde{p}$ defined as

$$E_{\tilde{p}}[f_i]=\sum_{x,y}\tilde{p}(x,y)f_i(x,y)=\sum_{x,y}\tilde{p}(x)\tilde{p}(y|x)f_i(x,y). \tag{5}$$

A nice property of the MaxEnt model with moment constraints (Berger et al., 1996) is that its solution is in the log-linear form of

$$p(y|x)=\frac{1}{Z_\lambda(x)}\exp\left(\sum_i\lambda_i f_i(x,y)\right), \tag{6}$$

where

$$Z_\lambda(x)=\sum_y\exp\left(\sum_i\lambda_i f_i(x,y)\right), \tag{7}$$

is a normalization constant to make sure $\sum_y p(y|x)=1$, and λ_i is the weight for the feature $f_i(x,y)$ and is chosen to maximize

$$\Psi(\lambda)=-\sum_x\tilde{p}(x)\log Z_\lambda(x)+\sum_i\lambda_i E_{\tilde{p}}[f_i]. \tag{8}$$

Since this dual problem is an unconstraint convex problem, many algorithms such as generalized iterative scaling (GIS) (Darroch and Ratcliff, 1972), gradient ascent and conjugate gradient (e.g., L-BFGS) (Nocedal, 1980), and RPROP (Riedmiller and Braun, 1993) can be used to find the solution. A comparison on the performance of different learning algorithms can be found in (Malouf, 2002 and Mahajan et al., 2006). Notice that applying the higher-order moment constraints in the MaxEnt model is equivalent to using higher-order statistics as features in the MaxEnt model with mean (i.e., first-order moment) constraint. The MaxEnt-MC model has been improved with regularization techniques (Chen and Rosenfeld, 1999, 2000; Goodman, 2004) and uncertain constraints (Kazama, 2004; Kazama and Tsujii, 2005) in the recent years.

2.2. Moment constraints on binary features and continuous features

The MaxEnt principle basically says one should not assume any additional structure or constraints other than those already imposed on the constraint set C. The appropriate selection of the constraints thus is crucial. In principle, we should include all the constraints that can be validated by (or reliably estimated from) the training set or prior knowledge.

With the binary features where $f_i(x,y)\in\{0,1\}$, the moment constraint described in Eq. (3) is a strong constraint since $E_p[f]=p(f=1)$. In other words, constraining the expected value implicitly constrains the probability distribution. However, the moment constraint is rather weak for continuous features. Constraining the expected value does not mean much to the continuous features because many different distributions can yield the same expected value. That is to say, much information carried in the training set is not used in the parameter estimation if solely moment constraints are used for the continuous features especially when the distribution of the features has multiple modes. This is the most important reason that the MaxEnt-MC model works well for binary features but not so well for non-binary features, especially the continuous features.

Let us illustrate this observation with an example. Consider a random process that generates 0 with probability 1 if $x\in\{1,3\}$, and generates 1 with probability 1 if $x\in\{2\}$, and assume that we have a training set with the empirical joint distributions

$$\begin{aligned}&\tilde{p}(1,0)=0.25,\quad \tilde{p}(1,1)=0,\\ &\tilde{p}(2,0)=0,\quad \tilde{p}(2,1)=0.5,\\ &\tilde{p}(3,0)=0.25,\quad \tilde{p}(3,1)=0,\end{aligned} \tag{9}$$

and features