PROCEEDINGS
of
The Twenty-seventh Annual
Conference
on
Information Sciences and Systems

Department of Electrical and Computer Engineering
The Johns Hopkins University
Baltimore, Maryland 21218

Maximum Partial Likelihood Estimation With Perceptrons*

M. Kemal Sönmez[1] *John S. Baras*[2†]
[1]Department of Mathematics and Institute for Systems Research
[2]Department of Electrical Engineering and Institute for Systems Research
University of Maryland, College Park, MD 20742

Abstract-We show the equivalence of two techniques of time series modelling/prediction; (i) perceptron learning of probability distribution of the truth value of a proposition from first order stochastic density approximations, (ii) Maximum Partial Likelihood (MPL) estimation of the parameters of a logistic regressive model for binary time series. This result provides large training set characteristics for the approximate Kullback-Leibler relative entropy learning scheme.

I. Introduction

The main motivation of this work is considering the sigmoid analog perceptron as a parametrised model of a probability distribution and formulating the learning problem as approximation to a desired conditional probability distribution. In [1,5], minimizing the Kullback-Leibler (K-L) relative entropy cost function on a multi-layer feed-forward network of analog units has been proposed for the problem of learning the correct probabilities of a set of propositions from analog input. In this work, we propose a time series modelling/prediction application for this scheme by using a first order approximation to the true probability density of the truth value of the proposition. We show that the learning scheme obtained by such an approach is equivalent to partial likelihood estimation of the parameters of a general logistic regressive model for a binary time series that takes into account stochastic time dependent covariates. Theoretical justification and large sample theory for this type of estimation have recently been developed by Slud and Kedem [2,4,7]. Therefore, we put the probability distribution learning from first order stochastic approximations in a statistical framework by using their results.

This result has relevance in signal processing with neural networks domain, since a neural estimation technique for dependent time series and its large sample properties are given, and in statistics, since an *analog* implementation for partial likelihood estimation of logistic models is pointed out.

The organization of the paper is as follows: In Section II, we define the time series prediction problem that we address. In Section III, we present a brief introduction to partial likelihood first introduced by Cox [3] and subsequently developed by Wong [6]. Section IV reviews the partial likelihood estimation of logistic regression model parameters framework of Slud and Kedem for the binary time series prediction problem. Relative entropy learning on an analog perceptron for the same binary time series problem is introduced in Section V. Equivalence of the two techniques and its implications are discussed in Section VI. We demonstrate the $-\ln PL = ARE$ surface and a stochastic gradient algorithm's behavior for an example process in Section VII. Finally, we conclude with a discussion of the use of more complex networks in this learning problem in Section VIII.

II. Task

We are interested in the prediction of the future truth values of a proposition from past information

*Research supported in part by NSF Grant NSFD CDR 8803012, through the Engineering Research Center Program and by a grant from Texas Instruments.

†Martin Marietta Chair in Systems Engineering

about the proposition and past and present information about some auxiliary stochastic analog time series referred to as *covariates* [4,7]. The covariates may consist of (functions of) other time series which are believed to affect how the proposition turns out and/or past truth values of the proposition. For example, if the proposition is that an autoregressive process exceeds a given threshold, the covariates are simply the past values of the process. If the proposition is that a certain component will fail, the covariates may consist of (functions of) operating conditions or periodic measurements on the component, etc. Define a binary time series $\{X_t\}$ as

$$X_t \equiv \begin{cases} 1, & \text{if proposition at t is true;} \\ 0, & \text{if proposition at t is false,} \end{cases}$$

and denote the covariate process with a column vector of time series, $\{\mathbf{Z}_t\}$. The information available about the covariates at time t, denoted by $\mathcal{F}_t$ is the σ-field $\sigma(X_t, X_{t-1}, ..., \mathbf{Z}_t, \mathbf{Z}_{t-1}, ...)$. It is clear that $\mathcal{F}_t \subset \mathcal{F}_{t+1}$. The prediction problem in this setting is the estimation of the probability that the proposition is true given all the past information which is given by

$$P(X_t = 1|\mathcal{F}_{t-1}) = E[X_t|\mathcal{F}_{t-1}].$$

This is essentially the same problem formulated in [4,7] as a level crossing problem, since any binary event can be regarded as a level crossing.

III. Partial Likelihood

Maximum likelihood estimation theory has mainly been developed for the i.i.d. observations where the full likelihood is a product of individual likelihoods. With dependent data, full likelihood is often hard or impossible to evaluate. Also in some cases where the full likelihood is available but complicated with nuisance parameters, the ML estimate may not exist (e.g. the contaminated Gaussian mixture).

Partial likelihood as introduced by Cox belongs to a class of approaches that attempt to use an appropriate factorization of the full likelihood. If the factorization has a factor depending only on useful parameters, then maximizing only this factor and giving up the information in the other factors that contain both useful and nuisance parameters leads to simplicity in analysis and robustness.

For the time series framework of Section II, the full likelihood of the joint process $\{X_t, \mathbf{Z}_t\}$ relative to some model parameters $\mathbf{w}$ and the data $\{X_t, \mathbf{Z}_t\}$ is given by

$$\begin{aligned} L(\mathbf{w}; X_1, ..., X_T) &= \prod_{t=1}^{T} p_{\mathbf{w}}(\mathbf{Z}_t|\mathcal{F}_{t-1}) \\ &\quad \times \prod_{t=1}^{T} p_{\mathbf{w}}(X_t|\sigma(\mathbf{Z}_t, \mathcal{F}_{t-1})) \quad (1) \end{aligned}$$

The second product in (1) is called the *partial likelihood* based on X [3].

The intuitive principles of partial likelihood are:
(i) Distribution of Z's should all depend essentially on nuisance parameters, i.e. Z's should not contain important information about parameters of interest.
(ii) Nuisance parameters should not occur in the partial likelihood.

Theoretical justification and ways of evaluating partial likelihood factorizations in terms of their large sample properties have been developed in [6]. For the specific particular binary time series problem in this work, Slud and Kedem have developed the large sample theory in this framework as described in the next section.

IV. Logistic Partial Likelihood

Logistic regression is a widely used model in expressing the relationship between the distribution of a random variable and a set of covariates. It has found wide application in statistics, health sciences and econometrics. The logistic regression model is remarkable in two respects from our perspective: (i) It has a neural network implementation, namely, the single-layer analog perceptron with a sigmoid response (Fig. 1). (ii) It allows a large sample theory to be developed for the MPL estimation [4,7]. The logistic model, in our case, is given by

$$\begin{aligned} p_t(\mathbf{w}) &\equiv P_{\mathbf{w}}(X_t = 1|\sigma(\mathbf{Z}_t, \mathcal{F}_{t-1})) \\ &= \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{Z}_{t-1})}. \end{aligned} \quad (2)$$

With this model the partial likelihood is conveniently written as

$$PL(\mathrm{w}; X_1, ..., X_T) = \prod_{t=1}^{T} p_t(\mathrm{w})^{X_t}(1-p_t(\mathrm{w}))^{1-X_t}. \quad (3)$$

The maximum partial likelihood estimate (MPLE), $\hat{\mathrm{w}}$ is defined as

$$\hat{\mathrm{w}} = \arg\{\sup_{\mathrm{w}} PL(\mathrm{w}; X_1, ..., X_T)\}. \quad (4)$$

We can use ln-PL instead of PL, therefore the estimation consists of maximizing the function

$$\ln PL(\mathrm{w}) = \sum_{t=1}^{T} [X_t \ln(p_t(\mathrm{w})) + (1-X_t)\ln(1-p_t(\mathrm{w}))]. \quad (5)$$

V. Relative Entropy Learning

The natural cost function in learning probability densities is the Kullback-Leibler relative entropy, the information theoretic distance between the desired probability density and the network output:

$$\begin{aligned} RE(\mathrm{w}) &= \sum_{t=1}^{T} p_t^{true} \ln\left(\frac{p_t^{true}}{p_t(\mathrm{w})}\right) \\ &+ \sum_{t=1}^{T}(1-p_t^{true})\ln\left(\frac{1-p_t^{true}}{1-p_t(\mathrm{w})}\right), \quad (6) \end{aligned}$$

where

$$p_t^{true} = P(X_t = 1|\mathcal{F}_{t-1}).$$

Since the true probabilities are not available for training, exact relative entropy can not be used. We need to train the network with some approximate estimate of the pdf. For an on-line application, first order approximation gives:

$$\begin{aligned} p_t^{true} &= P(X_t = 1|\mathcal{F}_{t-1}) \\ &= E[X_t|\mathcal{F}_{t-1}] \\ &\approx X_t. \quad (7) \end{aligned}$$

This results in an approximate cost function, Approximate Relative Entropy (ARE) :

$$\begin{aligned} ARE(\mathrm{w}) &= \sum_{t=1}^{T} X_t \ln\left(\frac{X_t}{p_t(\mathrm{w})}\right) \\ &+ \sum_{t=1}^{T}(1-X_t)\ln\left(\frac{1-X_t}{1-p_t(\mathrm{w})}\right) \\ &\approx RE(\mathrm{w}). \quad (8) \end{aligned}$$

Intuitively, averaged over a large training set, this basic scheme should supply the true probabilities. In fact, we can go one step further and observe that since $X_t \in \{0,1\}$,

$$\begin{aligned} ARE(\mathrm{w}) &= -\sum_{t=1}^{T} X_t \ln(p_t(\mathrm{w})) \\ &\quad -\sum_{t=1}^{T}(1-X_t)\ln(1-p_t(\mathrm{w})) \\ &= -\ln PL. \quad (9) \end{aligned}$$

We have thus shown that relative entropy learning with (9) is equivalent to MPL estimation with a logistic model developed in the previous section. Therefore, MPL estimators and network models have the same large sample properties. The extension of this result to multiple hypotheses is possible provided that the hypotheses are statistically independent. This condition is required to be able to write the partial likelihood as

$$PL(\mathrm{w}; \{X_t^i\}_{t=1,...,T}^{i=1,...,K}) = \prod_{i=1}^{K}\prod_{t=1}^{T} p_t^i(\mathrm{w})^{X_t^i}(1-p_t^i(\mathrm{w}))^{1-X_t^i}, \quad (10)$$

to give the cost function

$$\begin{aligned} ARE(\mathrm{w}) &= \sum_{i=1}^{K}\sum_{t=1}^{T} X_t^i \ln\left(\frac{X_t^i}{p_t^i(\mathrm{w})}\right) \\ &\quad +(1-X_t^i)\ln\left(\frac{1-X_t^i}{1-p_t^i(\mathrm{w})}\right). \quad (11) \end{aligned}$$

VI. Large Sample Properties

Theoretical justification and large sample properties of general partial likelihood estimators have been developed in [6]. In [4,7], the large sample properties of logistic MPLE's have been studied. The main result is, under some regularity conditions on the asymptotic behavior of time dependent covariates, as $T \to \infty$,

$$\sqrt{T}(\hat{\mathrm{w}} - \mathrm{w}) \longrightarrow N[0, \Lambda^{-1}(\mathrm{w})] \quad (12)$$

in law, where Λ is the information matrix per observation.

VII. Example: AR(2) Process

In this section, we present a numerical example: zero-crossing event of the the second order autoregressive process

$$Y_t = 0.5Y_{t-1} - 0.3Y_{t-2} + \epsilon_t, \tag{13}$$

where $\epsilon_t \sim \mathcal{N}(0, \pi^2/3)$ and

$$X_t \equiv I_{[(Y_t)\geq 0]}. \tag{14}$$

Here, the covariate vector is simply a finite window of past values of the process

$$\mathbf{Z}_{t-1} = [Y_{t-1} \;\; Y_{t-2}]^T, \tag{15}$$

and the prediction problem is

$$P(X_t = 1|\mathbf{Z}_{t-1}). \tag{16}$$

In [4,7], it is shown that for $\epsilon_t \sim f(x) = e^x/(1+e^x)$ which is very similar to $\mathcal{N}(0, \pi^2/3)$ (Fig. 2)

$$\mathbf{w} = [0.5, -0.3]^T. \tag{17}$$

We show the process for $T = 100$ samples and the $-\ln PL = ARE$ surface in Figs. 3 and 4 respectively. Then, we depict the path of the stochastic gradient descent learning algorithm which is simply

$$\begin{aligned} \mathbf{w}_n &= \mathbf{w}_{n-1} - \alpha \nabla_{\mathbf{W}} ARE(\mathbf{w}) \\ &= \mathbf{w}_{n-1} - \alpha \sum_{t=1}^{T} (X_t - p_t(\mathbf{w}))\mathbf{Z}_{t-1}. \end{aligned} \tag{18}$$

The incremental version of this algorithm is particularly suitable for on-line adaptation since the data may be processed as they arrive.

VIII. Discussion

In summary, we have extended maximum likelihood estimation by minimizing the Kullback-Leibler cost function on a single-layer analog perceptron to dependent time series applications by using partial likelihood and shown that a large sample theory already exists. For the prediction to be successful, careful selection of (functions of) covariates is very important, since the auxiliary data must contain enough information to make an inference. It is conceivable that a multilayer perceptron may help the selection by training with backpropagation to generate appropriate functions of input data. This way, the preceding layers may preprocess the raw input data to present a vector of covariates to the final layer which will allow better inference. The nice large sample characteristics of logistic partial likelihood will however no longer be applicable since the model is not even unique. One remedy can be to stop training on the hidden layer after a point and return to the logistic model with the transformed covariates. Provided the transformed covariates satisfy the asymptotic behavior conditions, the large sample theory can be recovered.

Acknowledgement: The authors are grateful to Benjamin Kedem for providing a chapter from his forthcoming book and useful discussions.

References

[1] Hopfield, J.J. "Learning algorithms and probability distributions in feed-forward and feed-back networks," *Proc. Natl. Acad. Sci. USA* vol. 84, pp. 8429-8433(1987).

[2] Chiu, L.S. and Kedem, B. "Estimating the Exceedance Probability of Rain Rate by Logistic Regression," *Journal of Geophysical Research* vol. 95, pp. 2217-2227(1990).

[3] Cox, D.R. "Partial Likelihood," *Biometrika* vol. 62, no. 2, pp. 269-276(1975).

[4] Slud, E. and Kedem, B. "Partial likelihood analysis of logistic regression and autoregression," Report TR91-47/ES/BK, Department of Mathematics, University of Maryland, College Park, 1991.

[5] Hertz, J., Krogh, A., Palmer, R.G.," *Introduction to the Theory of Neural Computation*, Addison-Wesley, Redwood City, CA, 1991.

[6] Wong, W.H.. "Theory of Partial Likelihood," *The Annals of Statistics* vol. 14, no. 1, pp. 88-123(1986).

[7] Kedem, B.," *Time Series Analysis by Higher Order Crossings*, Forthcoming from IEEE Press.

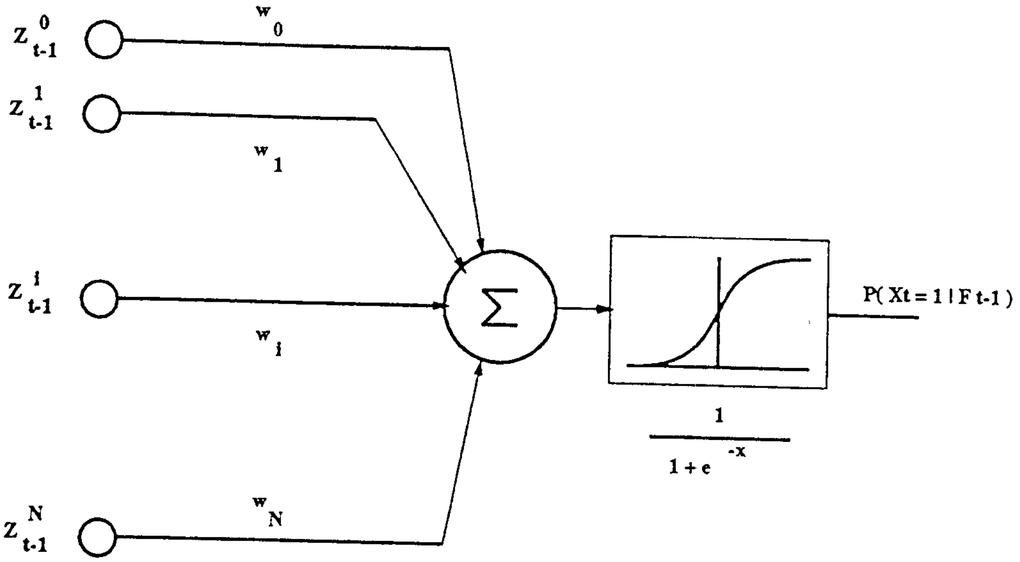

Figure 1: Perceptron as a model for a probability distribution

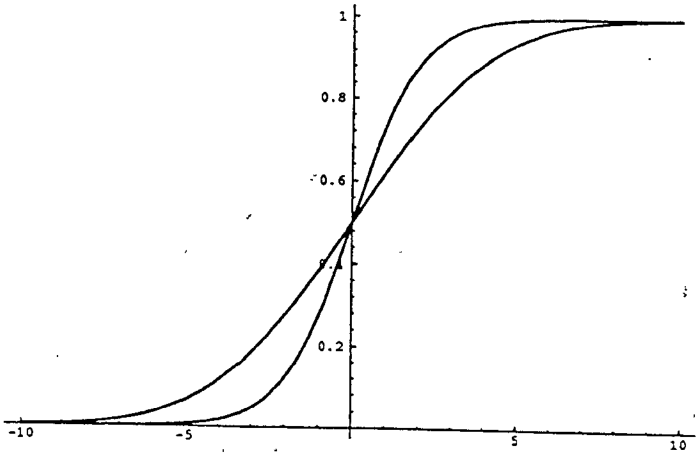

Figure 2: Logistic and normal cdf's

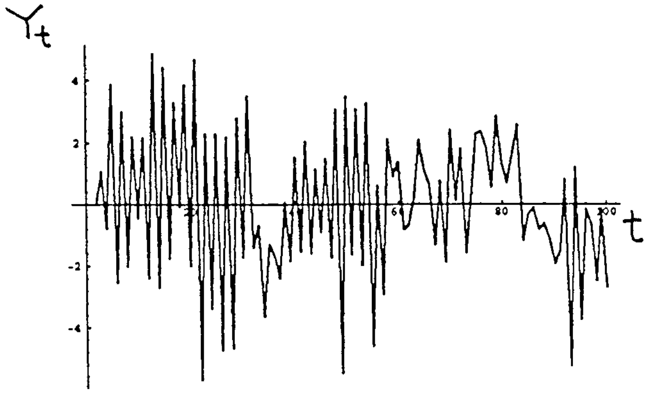

Figure 3: AR(2) process, T=100

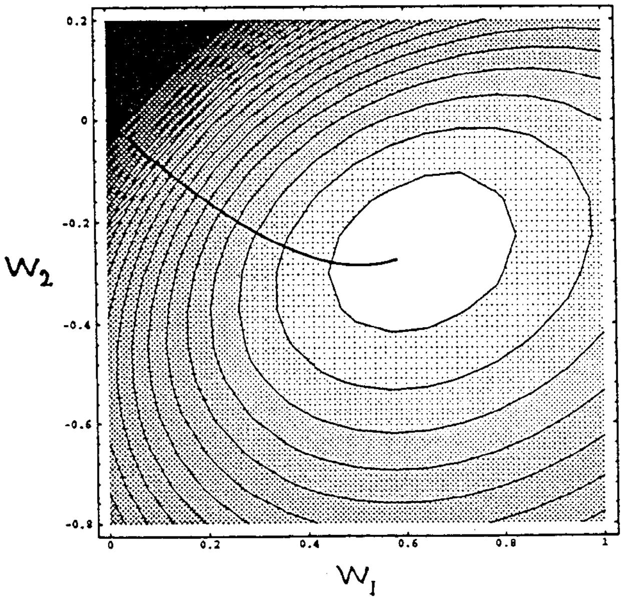

Figure 4: $ARE = -\ln PL$ surface, the path of the stochastic gradient.