JB 93-04

Entitled:

"Time Series Modeling by Perceptrons:
A Likelihood Approach"

Authors:

(with M.K. Sonmez)

1993 International Neural Network Society Annual Meeting

Portland, Oregon

July 1993

Time Series Modeling by Perceptrons: A Likelihood Approach [1]

M. Kemal Sönmez[*] *John S. Baras*[†2]

[*]Department of Mathematics and Institute for Systems Research
[†]Department of Electrical Engineering and Institute for Systems Research
University of Maryland, College Park, MD 20742

Abstract

We consider neural network learning problems in which the objective is to learn the relationship between the inputs and the probability distribution of a proposition. We regard successive truth values of the proposition as a dependent binary time series whose instantaneous probability of truth is a function of the past behavior of the joint process between the analog inputs and binary output truth values. In this context, we *identify* the gradient descent learning algorithm using the Kullback-Leibler relative entropy cost function on a perceptron *with* a Maximum Partial Likelihood (MPL) estimator of the perceptron model for the probability of a binary event in terms of its covariates. The implications of this result are: (i) The neural network models obtained by relative entropy learning are shown to have the nice large sample (i.e. training set size) characteristics of MPL estimates: consistency and asymptotic normality. (ii) An important and widely used statistical inference technique, logistic regression, can efficiently be implemented on analog perceptrons for time series modelling and prediction.

I. Introduction

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 hypotheses from analog input [1,5]. In this work, we propose a time series modelling/prediction application for this scheme by using it to implement an estimation technique whose numerical and large sample properties have recently been studied [2,4]. The results reported below have relevance (i) in signal processing with neural networks domain, since a neural estimation technique for dependent time series and its large sample properties are given, (ii) in supervised learning, since probability distribution learning with the K-L relative entropy cost function is put in a statistical perspective, and (iii) in statistics, since an *analog* implementation for partial likelihood estimation of logistic models is pointed out.

II. Task

We are interested in the prediction of the future truth values of a proposition from past information about the proposition and past and present information about some auxiliary stochastic analog

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

[2] Martin Marietta Chair in Systems Engineering

time series referred to as *covariates* [4]. 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

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

III. Maximum Partial Likelihood Estimation

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

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

The second product in (1) is called the *partial likelihood* based on X [3]. In cases where the full likelihood is unknown or hard to evaluate, such as when dealing with dependent data, partial likelihood may still be used to obtain ample statistical inference. Also in cases where a maximum likelihood estimate does not exist due to nuisance parameters, partial likelihood may be used to obtain estimates that preserve most of the information. It is a generalization of both marginal likelihood and conditional 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. (ii) It allows a large sample theory to be developed for the MPL estimation. The logistic model, in our case, is given by

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

With this model the partial likelihood is conveniently written as

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

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

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

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

$$G(\mathbf{w}) = \sum_{t=1}^{T} \left[X_t \ln(p_t(\mathbf{w})) + (1 - X_t) \ln(1 - p_t(\mathbf{w}))\right]. \tag{5}$$

IV. Relative Entropy Learning

The main result of this development is that the maximization of $G(\mathbf{w})$ is equivalent to minimization of the K-L relative entropy cost function

$$E(\mathbf{w}) = \sum_{t=1}^{T} \left[X_t \ln\left(\frac{X_t}{p_t(\mathbf{w})}\right) + (1 - X_t) \ln\left(\frac{1 - X_t}{1 - p_t(\mathbf{w})}\right)\right]. \tag{6}$$

when $X_t \in \{0, 1\}$, i.e. when we use the following approximation in learning

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

This basic scheme should implicitly supply the probabilities when averaged over a large training set, as pointed out in [1]. It is well-known that for a gradient descent algorithm we obtain

$$\nabla_{\mathbf{w}} E(\mathbf{w}) = \sum_{t=1}^{N} (X_t - p_t(\mathbf{w}))\mathbf{Z}_{t-1}. \tag{7}$$

We have shown that relative entropy learning with (6) is equivalent to MPL estimation with a logistic model developed in the previous section. Therefore, the large sample properties for MPL estimators also hold for network models. 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

$$PL(\mathbf{w}; \{X_t^i\}_{t=1,...,T}^{i=1,...,K}) = \prod_{i=1}^{K} \prod_{t=1}^{T} p_t^i(\mathbf{w})^{X_t^i} (1 - p_t^i(\mathbf{w}))^{1 - X_t^i}. \tag{8}$$

to give the cost function

$$E(\mathbf{w}) = \sum_{i=1}^{K} \sum_{t=1}^{T} \left[X_t^i \ln\left(\frac{X_t^i}{p_t^i(\mathbf{w})}\right) + (1 - X_t^i) \ln\left(\frac{1 - X_t^i}{1 - p_t^i(\mathbf{w})}\right)\right]. \tag{9}$$

V. Large Sample Properties

The large sample properties of MPLE's have been studied in [4]. The main result is, under some regularity conditions, as $T \to \infty$,

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

in law, where $\mathbf{\Lambda}$ is the information matrix per observation defined by

$$\mathbf{\Lambda}(\mathbf{w}) = \int \frac{\exp(\mathbf{w}^T\mathbf{z})}{(1+\exp(\mathbf{w}^T\mathbf{z}))^2}\mathbf{z}\mathbf{z}^T \nu(d\mathbf{z}).$$

VI. Discussion

In summary, we have extended the 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 multi-layer 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.

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.