A NEURAL NETWORK FRAMEWORK FOR IMPLEMENTING THE BAYESIAN LEARNING

Luminita State
University of Pitesti, Caderea Bastiliei #45, Bucharest #1, Romania

Catalina Cocianu, Viorica Stefanescu
Academy of Economic Studies, Calea Dorobantilor 15-17, Bucharest #1, Romania

Vlamos Panayiotis
Hellenic Open University, Greece

Keywords: Neural Networks, Competitive Learning, Hidden Markov Models, Pattern Recognition, Bayesian Learning, Weighting Processes, Markov Chains.

Abstract: The research reported in the paper aims the development of a suitable neural architecture for implementing the Bayesian procedure for solving pattern recognition problems. The proposed neural system is based on an inhibitive competition installed among the hidden neurons of the computation layer. The local memories of the hidden neurons are computed adaptively according to an estimation model of the parameters of the Bayesian classifier. Also, the paper reports a series of qualitative attempts in analyzing the behavior of a new learning procedure of the parameters an HMM by modeling different types of stochastic dependencies on the space of states corresponding to the underlying finite automaton. The approach aims the development of some new methods in processing image and speech signals in solving pattern recognition problems. Basically, the attempts are stated in terms of weighting processes and deterministic/non deterministic Bayesian procedures. The aims were mainly to derive asymptotical conclusions concerning the performance of the proposed estimation techniques in approximating the ideal Bayesian procedure. The proposed methodology adopts the standard assumptions on the conditional independence properties of the involved stochastic processes.

1 HMM IN BAYESIAN LEARNING

Stochastic models represent a very promising approach to temporal pattern recognition. An important class of the stochastic models is based on Markovian state transition, two of the typical examples being the Markov model (MM) and the Hidden Markov Model (HMM).

The latent structure of observable phenomenon is modeled in terms of a finite automaton *Q*, the observable variable being thought as the output produced by the states of *Q*. Both evolutions, in the spaces of non observable as well as in the space of observable variables, are assumed to be governed by probabilistic laws.

In the sequel, we denote by $(\Lambda_n)_{n\geq 0}$ the stochastic process describing the hidden evolution and by $(X_n)_{n\geq 0}$ the stochastic process corresponding to the observable evolution.

Let *Q* be the set of states of the underlying finite automaton; $|Q| = m$. We denote by τ_n the probability distribution on *Q* at the moment *n*. Let $(\Omega, \Im, P)$ be a probability space, $(\aleph, C, \sigma)$ be a measure space, where σ is a σ -finite measure. The output of each state $q \in Q$ is represented by the random element $X : \Omega \to \aleph$ of density function $f_q(.)$. Let ξ be the *apriori* probability distribution on Q. We assume that $\forall q \in Q, \xi(q) \neq 0$. The

State L., Cocianu C., Stefanescu V. and Panayiotis V. (2004).
A NEURAL NETWORK FRAMEWORK FOR IMPLEMENTING THE BAYESIAN LEARNING.
In *Proceedings of the First International Conference on Informatics in Control, Automation and Robotics*, pages 328-331
DOI: 10.5220/0001127503280331

conclusions on the hidden evolution are derived using the Bayesian procedure when the *apriori* probability distribution ξ and the set of density functions $(f_{n,q}, q \in Q)$ are known.

Let $L : Q \times Q \to [0,\infty)$ be a risk function. The outputs of the automaton are represented by the sequence of random elements $(X_n)_{n\geq 0}$, where the output at the moment *n*, X_n is distributed $\rho(q_n)$ if it was emitted by the state q_n. Let $R = \left\{ t / t \in \left([0,1]^Q\right)^{\aleph} \right\}$ be the set of random decision procedures, where, for any $t \in R, q \in Q, x \in \aleph$, $t(x)(q)$ is the probability of deciding that the output *x* is produced by the state *q*.

For any $t \in R$ we denote the expected risk by,

$$R(\xi, t, f) = \sum_{q\in Q}\sum_{\bar{q}\in Q}\int_{\aleph} \xi(q)L(q,\bar{q})t_{\bar{q}}(x)f_q(x)\sigma(dx)$$

The Bayesian decision procedure $\tilde{t} \in R$ assures the minimum risk that is,

$$R(\xi, \tilde{t}, f) = \inf_{t\in R} R(\xi, t, f) \triangleq \Phi(\xi, f)$$

and it is given by,

$$(1)\ \tilde{t}_{\bar{q}}(x) = \begin{cases} 1, & T(\bar{q},x) < \min\limits_{q^*\in Q\setminus\{\bar{q}\}} T(q^*,x) \\ 0, & T(\bar{q},x) > \min\limits_{q^*\in Q\setminus\{\bar{q}\}} T(q^*,x) \\ \alpha_{\bar{q}}, & T(\bar{q},x) = \min\limits_{q^*\in Q\setminus\{\bar{q}\}} T(q^*,x) \end{cases}, \text{ where}$$

$$(2)\quad T(\bar{q},x) = \sum_{q\in Q}\xi(q)L(q,\bar{q})f_q(x),\ \sum_{\bar{q}\in A}\alpha_{\bar{q}} = 1,$$

$\forall \bar{q} \in A, \alpha_{\bar{q}} \geq 0$ and

$$A = \left\{ \bar{q} / \bar{q} \in Q, T(\bar{q},x) = \min_{q^*\in Q\setminus\{\bar{q}\}} T(q^*,x) \right\}.$$

The true evolution in the space *Q* of non observable variables is governed by probabilistic laws, $(\tau_n)_{n\geq 0}$, where τ_n represents the probability distribution on *Q* at the moment *n*.

Let $(u_n)_{n\geq 0}$ be a sequence of subjective utilities assigned to the states of the automaton; $\forall n \geq 0, u_n : Q \to [0,\infty)$. We assume that, for any $n \geq 1, \sum_{q\in Q} u_n(q) \neq 0$. For any $n \geq 0$ and $q \in Q$, $u_n(q)$ stands for the subjective utility assigned to the state *q* at the moment *n*. Typically, $u_n(q)$ can be taken as the relative emitting frequency of the state q during the time interval $[0,n]$.

Let $(g_n)_{n\geq 1}$ be a sequence of measurable functions, $g_n : \aleph \times \aleph \to [0,\infty)$, $\forall n \geq 1$, a Parzen-like basis of asymptotically unbiased estimates of the system of density functions $(f_q, q \in Q)$ satisfying a series of convenable regularity assumptions. Our method is a supervised technique based on the learning sequence $S = ((\Lambda_n, X_n) / n \geq 1)$, where the true probability distribution τ_n is approximated by a weighting process $(\xi_n(q), q \in Q)_{n\geq 0}$ defined by $\xi_n(q) = \dfrac{\xi(q)u_n(q)}{\sum_{\bar{q}\in Q}\xi(\bar{q})u_n(\bar{q})}$ representing the guess that *q* is the emitting state at the moment *n*. The decision procedure $\tilde{t}_n^*$ is defined by (1) in terms of $\xi_n(q)$ and $f_{n,q}(x) = \dfrac{1}{n\xi_n(q)}\sum_{j=1}^{n}\delta(\Lambda_j, q)g_n(x, X_j)$, where $\delta(q,\bar{q}) = \begin{cases} 1, q = \bar{q} \\ 0, q \neq \bar{q} \end{cases}$. The criterion function $T(\bar{q},x)$ given by (2) is replaced by $(3)\ T(\bar{q},x) = \sum_{q\in Q}\xi_n(q)L(q,\bar{q})f_{n,q}(x)$.

2 THEORETICAL RESULTS SUPPORTING THE QUALITATIVE ANALYSIS OF THE BEHAVIOR OF THE LEARNING SCHEME

Let $R_\xi(\tilde{t}_n^*) = E(R(\xi, \tilde{t}_n^*, f))$ be the expected risk corresponding to the random decision procedure $\tilde{t}_n^*$ when ξ is the true probability distribution on *Q* and $f = (f_{n,q}, q \in Q)$ is the set of output density functions.

Theorem 1. (State, 2002) Let $(g_n)_{n\geq 0}$ be a sequence of measurable functions such that the assumptions A_1, A_2, A_3, A_4 hold, where,

(A_4) for any $k \geq 1$, $q \in Q, x \in \aleph$,

$(A_4')\ E_q(g_k(x,X)) = f_q(x)$.

If $S = ((\Lambda_n, X_n) / n \geq 1)$ is a learning sequence such that the random elements $(\Lambda_n, X_n), n \geq 1$ are independent, Λ_n is distributed ξ and X_n is distributed f_q if $\Lambda_n = q$, then, for the Parzen-like basis $(g_n)_{n\geq 0}$, $\lim\limits_{n\to\infty} R_\xi(\tilde{t}_n^*) = \Phi(\xi, f)$.

Theorem 2. (State, 2002) Let $S=((\Lambda_n, X_n)/\, n\geq 1)$ be a learning sequence such that the random elements $(\Lambda_n, X_n), n\geq 1$ are independent, Λ_n is distributed τ_n and X_n is distributed f_q if $\Lambda_n = q$. If for the sequence $(g_n)_{n\geq 0}$, the assumptions A_1, A_2, A_3, A_4 hold and, for any $q\in Q$, $\lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\tau_j(q)=\tau(q)$, then,

$$\lim_{n\to\infty} E\left(R\left(\tau,\tilde{t}_n^*, f\right)\right)=\Phi(\xi, f).$$

Theorem 3. (State, 2002) Assume that the conditions mentioned in theorem 2 hold. If, for any $q\in Q$, $\lim_{n\to\infty}\tau_n(q)=\tau(q)$, then,

$$\lim_{n\to\infty} E\left(R\left(\tau_n,\tilde{t}_n^*, f\right)\right)=\Phi(\tau, f).$$

Theorem 4. (State, 2002) Let $S=((\Lambda_n, X_n)/\, n\geq 1)$ be a learning sequence such that $(\Lambda_n, n\geq 1)$ is a Markov chain of stationary transition probabilities having an unique recurrent class Q'. If $(X_n), n\geq 1$ are independent and X_n is distributed f_q if $\Lambda_n = q$, then

$$\lim_{n\to\infty} E\left(R\left(\tau,\tilde{t}_n^*, f\right)\right)=\Phi(\tau, f),$$

where τ is the probability distributions of Λ_1.

If $(\Lambda_n, n\geq 1)$ is a Markov chain then $(\Lambda_n, X_n), n\geq 1$ is a Markov chain of stationary transition probabilities having an unique recurrent class $R'=\bigcup_{q\in Q'}\{(q,x)/\, x\in C_q\}$, where $C_q=\{x/\, x\in\aleph,\ f_q(x)\neq 0\}$.

3 NEURAL ARCHITECTURE FOR IMPLEMENTING THE BAYESIAN PROCEDURE $\tilde{t}_n^*$

We assume that $\aleph = R^d$. Then the neural architecture consists of the layers F_X, F_H of d and respectively $|Q|$ neurons. The neurons of the input layer F_X have no local memory, they distribute the corresponding inputs toward the neurons of the hidden layer F_H. Each neuron of F_H is assigned to one of the pattern classes from Q. For simplicity sake, we'll refere to each neuron of F_H by its corresponding pattern class. The local memory of each neuron $q\in F_H$ consists of $\xi_n(q)$ and the parameters needed to compute $f_{n,q}$. The activation function of the neuron $q\in F_H$ at the moment n is $h_{n,q}(x)=f_{n,q}(x)\xi_n(q)$. The layer F_H is fully connected, the connection from q to $\bar{q}$ is weighted by $(-L(q,\bar{q}))$. Consequently, the input $x=(x_1,...,x_d)$ applied to F_X induces the neural activations,

$$net(\bar{q},0)=-\sum_{q\in Q}\xi_n(q)L(q,\bar{q})f_{n,q}(x)=-T(\bar{q},x),\ \bar{q}\in F_H$$

The recognition task corresponds to the identification of the states $\bar{q}$ for which $T(\bar{q},x)$ is minimum. This task is solved by installing a discrete time competitive process among the neurons of F_H. Let $S_q(t)=f(net(q,t))$ be the output of the neuron $q\in F_H$ at the moment t, where the competition process starts at the moment 0 and the activation function f is given by $f(u)=\begin{cases}0, u\geq 0\\ u, u<0\end{cases}$. We denote by $S(t)=(S_q(t), q\in F_H)$ the state at the moment t. The initial state is $S(0)=(f(net(q,0)), q\in F_H)$.

The synaptic weights of the connections during the competition are, $w_{q,\bar{q}}=\begin{cases}1, q=\bar{q}\\ -\varepsilon, q\neq\bar{q}\end{cases}$, where $\varepsilon>0$ is a vigilance parameter.

The update of the state is performed synchronously, that is, for any $q\in F_H$,

$$net(q,t+1)=S_q(t)-\varepsilon\sum_{\bar{q}\neq q}S_{\bar{q}}(t)=$$
$$=(1+\varepsilon)S_q(t)-\varepsilon\sum_{\bar{q}\in F_H}S_{\bar{q}}(t)$$
$$S_q(t+1)=f(net(q,t+1)).$$

The conclusions concerning the behavior of the competition in the space of states stem from the following arguments. Note that $S_q(t)\leq 0$, for any $t\geq 0$ and $q\in F_H$.

1. If $S_q(t)=0$, then $h_q(t+1)\geq 0$, hence $S_q(t+1)=0$. Moreover, for any $t'\geq t, S_q(t')=0$.

2. Assume that for some $q,q'\in F_H$, $t\geq 0$, $S_{q'}(t)=S_q(t)<0$. Then for any $t'\geq t$, $S_{q'}(t')=S_q(t')\leq 0$.

3. Assume that for some $q,q'\in F_H$, $t\geq 0$, $S_{q'}(t)<S_q(t)<0$. Then, $S_{q'}(t+1)\leq S_q(t+1)$. Moreover, for any $t'\geq t$, $S_{q'}(t')\leq S_q(t')$.

Using some of the previous arguments, we get that there exists $t(q') \geq 0$ such that $S_{q'}(t) = 0$ for any $t \geq t(q')$.

4. Assume that for $q, q' \in F_H$,

$0 < T(q,x) < T(q',x)$,

Then, for any $t \geq 0$, $S_{q'}(t) \leq S_q(t) \leq 0$ hence there exists $t(q') \geq 0$ such that $S_{q'}(t) = 0$ for any $t \geq t(q')$. Therefore, the competition installed by the above mentioned process among the neurons of F_H determines that the outputs of all neurons q' that received values $T(q',x) > \min_{q \in F_H} T(q,x)$ are inhibited in a finite number of stages, that is there exists t_{fin} such that $S_{q'}(t_{fin}) \neq 0$ if and only if

$$T(q',x) = \min_{q \in F_H} T(q,x).$$

Also, for any $q', q'' \in F_H$ such that

$$T(q',x) = T(q'',x) = \min_{q \in F_H} T(q,x),$$

$$S_{q'}(t_{fin}) = S_{q''}(t_{fin}) \neq 0$$

and for any $t \geq 0$, $S_{q'}(t) = S_{q''}(t)$.

The local memories of the hidden neurons are determined in a supervised way by adaptive learning algorithms using a learning sequence $S = ((\Lambda_n, X_n) / n \geq 1)$. The recurrent relations for $f_{n,q}, \xi_n(q), n \geq 1, q \in F_H$ are derived in terms of the particular relationships of $(u_n(q)), (g_n(x,y))$. For instance, if $u_n(q) = \frac{\sum_{j=1}^{n} \delta(\Lambda_j, q)}{n}$ and $g_n(x,y) = \delta(x,y)$ then we get the following relations. Let $\Lambda_{n+1} = q_{n+1}$. Then

$$u_{n+1}(q) = \begin{cases} \frac{n}{n+1} u_n(q), q \neq q_{n+1} \\ \frac{n u_n(q) + 1}{n+1}, q = q_{n+1} \end{cases},$$

therefore $u_{n+1}(q) = \frac{n}{n+1} u_n(q) + \frac{1}{n+1} \delta(q, q_{n+1})$

For $q \neq q_{n+1}$ we get,

$$\xi_{n+1}(q) = \frac{n \xi(q) u_n(q)}{n \sum_{q' \in F_H} \xi(q') u_n(q') + \xi(q_{n+1})}.$$

If $\xi(q) = 0$ or $u_n(q) = 0$, then $\xi_n(q) = \xi_{n+1}(q) = 0$. If $\xi(q) \neq 0, u_n(q) \neq 0$, then, denoting by $P_n(q) = \frac{1}{\xi_n(q)}$, we get

$$P_{n+1}(q) = P_n(q) + \frac{\xi(q_{n+1})}{n \xi(q) u_n(q)}.$$

For $q = q_{n+1}$ we get

$$\xi_{n+1}(q_{n+1}) = \frac{\xi(q_{n+1})(1 + n u_n(q_{n+1}))}{n \sum_{q' \in F_H} \xi(q') u_n(q') + \xi(q_{n+1})},$$

that is

$$P_{n+1}(q_{n+1}) = P_n(q_{n+1}) \frac{n u_n(q_{n+1})}{1 + n u_n(q_{n+1})} + \frac{1}{1 + n u_n(q_{n+1})}$$

Since

$$(n+1) f_{n+1,q}(x) \xi_{n+1}(q) = \delta(\Lambda_{n+1}, q) \delta(x, X_{n+1}) + n \xi_n(q) f_{n,q}(x)$$

we get, for $q \neq q_{n+1}$,

$$f_{n+1,q}(x) = \frac{n}{n+1} f_{n,q}(x) \left[1 + \frac{\xi(q_{n+1}) \xi_n(q)}{n \xi(q) u_n(q)} \right]$$

and respectively,

$$f_{n+1,q_{n+1}}(x) = \frac{\delta(x, X_{n+1}) + n \xi_n(q_{n+1}) f_{n,q_{n+1}}(x)}{(n+1) \xi_{n+1}(q_{n+1})}.$$

REFERENCES

Bishop, C., 1996, *Neural Networks for Pattern Recognition*; Oxford University Press

Devroye, L., Gyorfi, L., Lugosi, G., 1996.; *A Probabilistic Theory of Pattern Recognition,* Springer Verlag

Fukunaga, K., 1990, *Introduction to Statistical Pattern Recognition,* Academic Press

Lampinen, J., Vehtari, A, 2001 Bayesian Approach for Neural Networks: Review and Case Studies In *Neural Networks*, Vol. 14

State, L., Cocianu, C., 2001, Information Based Algorithms in Signal Processing In *Proceedings of SYNASC'2001* (The 3rd International Workshop on Symbolic and Numeric Algorithms for Scientific Computation), Timişoara, 3-5 octombrie 2001.

State, L., Cocianu, C., Vlamos, P., 2002, Nonparametric Approach to Learning the Bayesian Procedure for Hidden Markov Models In *Proceedings of SCI2002*, Orlando, USA, July 14-18, 2002

Stewart, W., 1994, *Introduction to the Numerical Solutions of Markov Chains*; Princeton University Press