ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 377-382

Bayesian decision theory on three layered neural networks

Yoshifusa Ito[1], Cidambi Srinivasan[2]
[1]Department of Information and Policy Studies
Aichi-Gakuin University, Aichi-ken, 470-0195 Japan
ito@psis.aici-gakuin.ac.jp
[2]Department of Statistics, Patterson office tower
University of Kentucky, Lexington, Kentucky 40506, USA
srini@ms.uky.edu

Abstract. We treat the Bayesian decision problem, mainly the two-category case. A three layered neural network, having a logistic output unit and a small number of hidden layer units, can approximate the a posteriori probability in L^2-norm, without knowing the type of the probability distribution before learning, if the log ratio of the a posteriori probabilities is a polynomial of low degree as in the case of most familiar probability distributions. This is because the log ratio itself can be well approximated by a linear sum of outputs of the hidden layer units in L^2-norm.

1. Introduction

In the case of the two-category Bayesian classification, the log ratio g of the a posteriori probabilities (probability densities)

$$g(x) = \log P(\omega_1|x) - \log P(\omega_2|x) \tag{1}$$

and its monotone functions can be used as Bayesian discriminant functions [1]. Recently Funahashi (1998) has treated a special case of the two-category classification that the state-conditioned probabilities are normal. He proved that a three layered neural network having an output unit with the logistic activation function σ and at least $2d$ hidden layer units can approximate the contraction $\sigma(g)$ of the log ratio g in L^2-norm, showing that $\sigma(g) = P(\omega_1|x)$. Since σ is monotone increasing, $\sigma(g)$ can be sued as a Bayesian discriminant function.

In this paper, we prove that, in the case where the log ratio g of the a posteriori probabilities is linear (resp. linear or quadratic), the network having two (resp. $d+1$) hidden layer units can approximately realize the Bayesian discriminant function, g or $\sigma(g)$, for the two-category classification. Hence, the number $2d$ obtained by Funahashi is decreased to $d+1$. Moreover, the Bayesian discriminant function g itself can be approximated in the L^2-norm in our theorem. If the learning is not trapped at a local minimum, the network can realize the approximation without knowing the type of the probability

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 377-382

distribution. The tools are an approximation theorem in [3] and the result in [4]. We also discuss on the extension of our result to multicategory cases.

2. Bayesian decision theory and neural networks

The basic notation in this paper follows [1]. We use a three layered neural network having c output units and N hidden layer units, denoting the outputs by $F_w(\omega_i|x)$. We expect that $F_w(\omega_i|x)$ approximate the a posteriori probabilities $P(\omega_i|x)$ respectively. Let φ and ϕ be the activation functions of the output and the hidden layer units. Then,

$$F_w(\omega_i|x) = \varphi(\sum_{j=1}^{N} c_{ij}\phi(w_{ij}\cdot x + t_{ij}) + t_{i0}), \quad i = 1, ..., c, \tag{2}$$

where w is the connection weight and ω_i stand for the categories.

We summarize here the results described in [2] and [4], which are ingredients of this paper.

2.1 Theory by Ruck et al.

Let $(x, \omega) \in \mathbf{R}^d \times \Omega$, $\Omega = \{\omega_1, ..., \omega_c\}$ be teacher signals and let $\xi_i(\omega) = 1$ for $\omega = \omega_i$ and $\xi_i(\omega) = 0$ for $\omega \neq \omega_i$. Then, the mean square deviation of the outputs from $\xi_i(\omega)$ is

$$E(w) = \int_{\mathbf{R}^n} \sum_{i=1}^{c}\sum_{j=1}^{c} (F_w(\omega_i|x) - \xi_i(\omega_j))^2 p(\omega_j|x)p(x)dx$$

$$= \int_{\mathbf{R}^n} \sum_{i=1}^{c} \left\{((F_w(\omega_i|x) - 1))^2 p(\omega_i|x) + F(\omega_i x)^2(1 - p(\omega_i|x))\right\} p(x)dx$$

$$= e^2(w) + \int_{\mathbf{R}^n} \sum_{i=1}^{c} P(\omega_i|x)(1 - P(\omega_i|x))p(x)dx, \tag{3}$$

where

$$e^2(w) = \int_{\mathbf{R}^n} \sum_{i=1}^{c} ((F_w(\omega_i|x) - P(\omega_i|x))^2 p(x)dx$$

[4]. Since the second term on the most right hand side of (3) is independent of w, minimization of $E(w)$ by modifying w implies that of the first term.

Let $\{(x^{(j)}, \omega^{(j)})\}_{j=1}^{\infty}$ be a sequence of independent teacher signals with probability distribution $p(x, \omega)$. Then,

$$E^{(n)}(w) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{c} (F_w(\omega_i|x^{(j)}) - \xi_i(\omega^{(j)}))^2 \tag{4}$$

converges to $E(w)$ almost everywhere. If the gradient descent method is combined with (4), the outputs $F_{w^{(n)}}(\omega_i|x)$ of the network may converge to $P(\omega_i|x)$

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 377-382

respectively in the sense of $L^2(\mathbf{R}^d, p)$, where $w^{(n)}$ is the weight vector at the n-th step.

2.2 Funahashi's network

This result was applied by Funahashi to the case where $c = 2$ and $p(x, w_i)$, $i = 1, 2$, are normal ditributions [2]. In this case the decision function (1) is a quadratic form. His network has an output unit with the logistic activation function $\sigma(t) = (1+e^{-t})^{-1}$ and hidden layer units with a sigmoidal C^2-function ϕ as the activation function.

Noting that t^2 can be uniformly approximated by a linear sum of the form $\sum_{i=1}^{2} c_i\phi(\delta t + t_i) + t_0$ on any finite interval, he used a linear sum $\sum_{i=1}^{2d} c_i\phi(w_i \cdot x + t_i) + t_0$ to approximate a quadratic form. Since $\phi(w_i \cdot x + t_i)$ can be outputs of the hidden layer units, the output of the network is

$$\sigma(\bar{g}(x)) = \sigma(\sum_{i=1}^{2d} c_i\phi(w_i \cdot x + t_i) + t_0).$$

If g is the discriminant function defined by (1), $\sigma(g(x)) = p(\omega_1|x)$. Since σ is monotone increasing, $\sigma(g(x))$ is also a discriminant function. Since σ is a contraction, the uniform approximation of g by $\bar{g}$ on any compact set $\mathbf{K}$ implies approximation of $\sigma(g)$ by $\sigma(\bar{g})$ in $L^2(\mathbf{R}^d, p)$: $\int |\sigma(g(x)) - \sigma(\bar{g}(x))|^2 p(x)dx \leq \int_{\mathbf{K}} |g(x) - \bar{g}(x)|^2 p(x)dx + \int_{\mathbf{K}^c} p(x)dx \leq 2\varepsilon$.

3. Main Theorem

In this section, we prove that the log ratio g itself defined by (1) can be approximated in $L^2(\mathbf{R}^d, p)$. If g is linear the number of the hidden layer units can be only 2, and if g is a quadratic form the number can be as small as $d + 1$. The class of activation functions is sufficiently wide. For simplicity, we suppose that the probability distribution is continuous.

3.1. Approximation on R

The lemma below is a special case of the theorem obtained in [3].

Lemma 1. Let p be a probability measure on $\mathbf{R}$ such that $t \in L^2(\mathbf{R}, p)$ and let $h \in \mathbf{R}^d$. If $h^{(i)}$ is bounded for $i = 0, \cdots, k$, then, for any $\varepsilon > 0$, there is a constant δ for which

$$\|h'(0)\frac{d^i}{dt^i}t - \frac{d^i}{dt^i}\frac{1}{\delta}(h(\delta t) - h(0))\|_{L^p(\mathbf{R},p)} < \varepsilon \tag{5}$$

holds for $i = 0, \cdots, k$.

Furthermore, if $t^2 \in L^2(\mathbf{R}, p)$, then, for any $\varepsilon > 0$, there is a constant δ for which both (5) and

$$\|\frac{1}{2!}h''(0)\frac{d^i}{dt^i}t^2 - \frac{d^i}{dt^i}\frac{1}{\delta^2}(h(\delta t) - \delta h'(0)t - h(0))\|_{L^p(\mathbf{R},p)} < \varepsilon \tag{6}$$

hold for $i = 0, \cdots, k$.

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 377-382

Proof. We describe the proof of the latter half. Applying Maclaurin's theorem to $h(\delta t)$, we have that

$$\frac{1}{2!}h''(0)t^2 - \frac{1}{\delta^2}(h(\delta t) - \delta h'(0)t - h(0)) = \frac{1}{2!}(h''(0) - h''(\theta\delta t))t^2$$

for $i = 0$. By assumption, there is $M > 0$ for which the right hand side is bounded by Mt^2. Since $t^2 \in L^2(\mathbf{R}, p)$, there is a finite
interval $\mathbf{I}$ such that

$$\int_{\mathbf{I}^c} |\frac{1}{2!}(h''(0) - h''(\theta\delta t))|^2 t^4 dp(t) < \int_{\mathbf{I}^c} M^2 t^4 dp(t) < \frac{\varepsilon}{2}.$$

For the $\mathbf{I}$, there is $\delta > 0$ for which

$$\int_{\mathbf{I}} |\frac{1}{2!}(h''(0) - h''(\theta\delta t))|^2 t^4 dp(t) < \frac{\varepsilon}{2}$$

as g'' is continuous. Hence, we obtain that

$$\int |\frac{1}{2!}(h''(0) - h''(\theta\delta t))|^2 t^4 dp(t) = \int_{\mathbf{I}} + \int_{\mathbf{I}^c} < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon.$$

We have that, for $i = 1$,

$$h''(0)t - \frac{1}{\delta}(h'(\delta t) - h'(0)) = (h''(0) - h''(\theta\delta t))t = Mt,$$

and, for $i = 2$, $h''(0) - h''(\delta t) <$ M. Hence, similarly to the case i
$= 0$, we obtain (6) for $i = 1, 2$. For $i > 2$, we have that

$$|\frac{1}{2!}h''(0)\frac{di}{dt^i}t^2 - \frac{d^i}{dt^i}\frac{1}{\delta^2}(h(\delta t) - \delta h'(0)t - g(0))| = |\delta^{i-2}h^{(i)}(\delta t)| < \delta^{i-2}M.$$

The proof of (5) is similar and easier.

3.2. Approximation of the discriminant function

Let p be a probability measure defined on $\mathbf{R}^d$ and let $\mathbf{L}_w = \{tw| - \infty < t < \infty\}$ for $w \in \mathbf{S}^{d-1}$, where $\mathbf{S}^{d-1}$ is the unit sphere in $\mathbf{R}^d$. Denote by p_w the projection of p onto $\mathbf{L}_w$. A function f such that $h(x) = h(w \cdot x)$ is called a plane wave in the direction of $w \in \mathbf{S}^{d-1}$. For the h,

$$\int h(x)dp(x) = \int h(x)dp_w(x) = \int h(w \cdot x)dp_w(x).$$

The most right hand side of this equation can be regarded as an integral over the line $\mathbf{L}_w$. Hence, we can apply Lemma 1 to $h(w \cdot x)$.

We denote by Q_1 a linear function in x and by Q_2 a nonhomogeneous quadratic form.

Theorem 2. Let $h \in C^1(\mathbf{R})$ be a bounded nonconstant function, let p be a probability measure on $\mathbf{R}^d$ and let Q_1 be a linear function in $x \in \mathbf{R}^d$. If

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 377-382

$Q_1 \in L^2(\mathbf{R}^d, p)$, then, for any $\varepsilon > 0$, there are coefficients a_i, vectors $w_i \in \mathbf{S}^{d-1}$ and constants t_i, $i = 1,2$, such that

$$\|\bar{Q}_1 - Q_1\|_{L^2(\mathbf{R}^d,p)} < \varepsilon, \tag{7}$$

where

$$\bar{Q}_1(x) = \sum_{i=1}^{2} a_i h(w_i \cdot x + t_i) + t_0. \tag{8}$$

Let $g \in C^2(\mathbf{R})$ be a bounded nonconstant function, let p be a probability measure on $\mathbf{R}^d$ and let Q_2 be a quadratic form defined on $\mathbf{R}^d$. If $Q_2 \in L^2(\mathbf{R}^d, p)$, then, for any $\varepsilon > 0$, there are coefficients a_i, vectors $w_i \in \mathbf{S}^{d-1}$, $i = 1, \cdots, d+1$, and constants t_i, $i = 1, \cdots, d+1$, such that

$$\|\bar{Q}_2 - Q_2\|_{L^2(\mathbf{R}^d,p)} < \varepsilon, \tag{9}$$

where

$$\bar{Q}_2(x) = \sum_{i=1}^{d+1} a_i h(w_i \cdot x + t_i) + t_0. \tag{10}$$

Proof. We prove the latter half. By an appropriate nondegenerated linear transform of $(x_1, \cdots, x_d, 1)$ we have that $Q_2(x) = \sum_{i=1}^{d} c_i u_i^2 + c_0$ or $Q(x) = \sum_{i=1}^{d-1} c_i u_i^2 + c_n u_n$, where u_i are linear sums of x_i and a constant. This implies that, without loss of generality, we can decompose Q_2 into plane waves:

$$Q_2(x) = \sum_{i=1}^{d} c_i (w_i \cdot x + t_i)^2 + c_{d+1}(w_{d+1} \cdot x + t_{d+1}),$$

where some of c_i may be zero, $w_i \in \mathbf{S}^{d-1}$ and t_i are constants. We may suppose that $h'(0) \neq 0$ and $h''(0) \neq 0$. Otherwise, we can shift h so that this condition is satisfied. Then, by Lemmas 1, $(w_i \cdot x + t_i)^2$ can be approximated by

$$\frac{1}{h''(0)\delta^2} \left(h(\delta(w_i \cdot x + t_i) - \delta h'(0)(w_i \cdot x + t_i) - g(0)\right)$$

in the sense of $L^2(\mathbf{R}^d, p)$ with any accuracy. Consequently, $Q_2(x)$ can be approximated by

$$\sum_{i=1}^{d} a_i h(\delta(w_i \cdot x + t_i) + c_0(w_0 \cdot x + t_0)$$

in $L^2(\mathbf{R}^d, p)$. Again by Lemmas 1, $(w_0 \cdot x + t_0)$ can be approximated by

$$\frac{1}{h'(0)\delta} \left(h(\delta(w_0 \cdot x + t_0) - g(0)\right)$$

in $L^2(\mathbf{R}^d, p)$. Hence, $Q_2(x)$ is approximated by

$$\bar{Q}_2(x) = \sum_{i=1}^{d+1} a_i h(\delta(w_i \cdot x + t_i) + t_0$$

ESANN'2001 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 25-27 April 2001, D-Facto public., ISBN 2-930307-01-3, pp. 377-382

in $L^2(\mathbf{R}^d, p)$. This concludes the proof.

Note that $Q_k \in L^2(\mathbf{R}^d, p)$, $k = 1, 2$, is the prerequisite condition when we treat the approximation of Q_k in $L^2(\mathbf{R}^d, p)$. In this sense, we have obtained the slightest condition on Q_k and p. Lemma 1 implies that this result can be extended to derivatives.

4. Discussions

In the cases of many familiar probability distributions belonging to the exponential family, the log ratio $g(x)$ of a priori probabilities is a polynomial of low degree. For the binomial, polynomial or gamma distribution, $g(x)$ is a linear function, and for the normal distribution, $g(x)$ is a quadratic form. We can mention many such probability distributions. In these cases, the neural network having rather a small number of hidden layer units may approximately realize the Bayesian discriminant funtion without knowing the type of probability distribution before learning. If $g(x)$ is a polynomial of degree up to 2, the number can be $d + 1$. Accordingly, our result may be widely applied.

Theorem 2 implies that if a three layered neural network has a linear unit on the output layer, the output can approximate the discriminant function g in the sense of $L^2(\mathbf{R}^d, p)$. If the activation function of the unit is the logistic function σ, its output can approximate the a posteriori probability $\sigma(g(x)) = p(\omega_1|x)$. This approximation can also be realized in the sense of $L^2(\mathbf{R}^d, p)$, because $|\sigma(\bar{Q}_k(x) - \sigma(Q_k(x))| < |\bar{Q}_k(x) - Q_k(x)|$.

In the case of multicategory classification, $g_i(x) = \log P(\omega_i|x) - \log(1 - P(\omega_i|x))$ is not a polynomial even if the log likelihood ratio is a polynomial. Nevertheless the network may approximate the a posteriori probability, if it has sufficiently many hidden layer units. However, $g_{ij}(x) = \log P(\omega_i|x) - \log P(\omega_j|x)$ is a polynomial. Using this fact, we can construct a neural network which may output $\max\{P(\omega_i|x)|i = 1, ..., c\}$. The details will be described elsewhere with experimental results.

References

1. R.O. Duda, P.E. Hart: Pattern classification and scene analysis. John Wiley & Sons, (1973)
2. K. Funahashi: Multilayer neural networks and Bayes decision theory. Neural Networks, 11, 209-213 (1998)
3. Y. Ito: Simultaneous L^p-approximations of polynomials and derivatives on $\mathbf{R}^d$ and their applications to neural networks. (to appear)
4. M.D. Ruck, S. Rogers, M. Kabrisky, H. Oxley, B. Sutter: The multilayer perceptron as an approximator to a Bayes optimal discriminant function. IEEE Transactions on Neural Networks, I, 296-298 (1990)