

Estimating Conditional Distributions by Neural Networks

Kulczycki, P.; Schiøler, Henrik

Published in:
1998 IEEE World Congress on Computational Intelligence

Publication date:
1998

Document Version
Også kaldet Forlagets PDF

[Link to publication from Aalborg University](#)

Citation for published version (APA):
Kulczycki, P., & Schiøler, H. (1998). Estimating Conditional Distributions by Neural Networks. I *1998 IEEE World Congress on Computational Intelligence* (s. 1344-1349)

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy

If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Estimating Conditional Distributions by Neural Networks

Piotr Kulczycki

Cracow University of Technology, Faculty of Electrical Engineering, ul. Warszawska 24, PL-31-422 Cracow, Poland
Academy of Mining and Metallurgy, Faculty of Applied Mathematics, al. Mickiewicza 30, PL-30-059 Cracow, Poland

Henrik Schiøler

Aalborg University, Department of Control Engineering, Fredrik Bajers vej 7, DK-9220 Aalborg 0, Denmark

Abstract

Neural networks for estimating conditional distributions and their associated quantiles are investigated in this paper. A basic network structure is developed on the basis of kernel estimation theory, and consistency property is considered from a mild set of assumptions. A number of applications within statistics, decision theory and signal processing are suggested.

1. Introduction

Relationships between random variables are most often described by characteristic parameters such as mean vectors and covariance matrices or in extraordinary cases moments of higher order. When standard situations are considered, for example, if all variables are jointly Gaussian or when the conditional characteristics are linear or low degree polynomial functions, that approach is to be recommended. On the other hand when the situation is far from the above, more general methods should be considered. Statistical relationships are completely described by the joint distribution of all the variables in consideration, however in some cases it is appropriate to partition the considered random variables into two groups. One group of so called explanatory variables yielding information about or explaining the variables in the second group. From that point of view conditional distributions as well as their associated quantiles are the objects of interest. In some cases such conditional distributions are sufficiently precisely described by standard expressions with only a low number of characteristic parameters to be estimated statistically. In the remaining cases nonparametric methods including neural networks may prove to be useful, and the purpose of this paper is to develop a neural network applicable for estimating conditional distributions and quantiles in the general nonstandard situation.

Neural networks have in recent years developed into powerful tools for solving optimization problems within

e.g. classification, estimation and forecasting. For the majority of cases, the applied neural networks, from a statistical point of view, solve conditional estimation problems. The celebrated Back Propagation Error algorithm used for training Feed Forward Neural Networks is shown to be a special case of gradient optimization in the sense of mean squared error [1]. Feed Forward Neural Networks are analyzed in paper [2] for consistent estimation of conditional expectation functions, which optimize expected squared error. Optimal classification is concerned with the problem of classifying a set of objects, on the basis of feature measurements, while obtaining a minimal probability of misclassification. This problem is equivalent to conditional estimation, and it is shown in work [3] that Feed Forward Neural Networks estimate the optimal discriminating function, when trained with the Back Propagation Error Algorithm. In all of the above cases, some sort of optimization or training algorithm is applied adjusting initially random network parameters optimally w.r.t. average loss functions on a finite set of training data. A more constructive way to follow is indicated in paper [4], where a Probabilistic Neural Network for classification based on kernel estimators is investigated, as well as in works [5,6], in which a similar line is followed for proposing neural networks estimating conditional expectation functions. From a certain point of view, this strategy is the basis for suggesting a large class of different neural network architectures, including among others Localized Receptive Fields [7] and Counter Propagation Networks [8]. In this paper such a constructive strategy is pursued in order to design a Feed Forward Neural Network capable of estimating conditional distributions.

2. Mathematical Preliminaries

Consider a real random variable w with a distribution function F_w , and a number $p \in (0,1)$. Any real number q fulfilling

$$F_w(q) = p \quad (1)$$

is said to be a quantile of order p [9]. If the distribution function F_w is continuous and strictly monotonous, the quantile of order p is uniquely defined for all values p by equation (1). In general the p th order quantile $q(p)$ can be uniquely defined as follows

$$q(p) = \sup\{q \mid F_w(q) \leq p\}. \quad (2)$$

Consider two real random variables w and v defined on a common probability space with a joint distribution P_{wv} on the space \mathbb{R}^2 . Then the function $F_{w|v} : \mathbb{R}^2 \rightarrow [0, 1]$, exists ([10], section 33) such that

1. for every $v \in \mathbb{R}$, $F_{w|v}(\cdot, v)$ is a distribution function on the real space,
2. for every $A = (-\infty, d]$ and every measurable set $B \subset \mathbb{R}$

$$P_{wv}(A \times B) = \int_B F_{w|v}(d, v) dP_v(v). \quad (3)$$

The function $F_{w|v}$ is called the conditional distribution function of the random variable w with respect to v . In the case where the joint distribution P_{wv} has a density function h_{wv} , a conditional density function $h_{w|v}$ is given as

$$h_{w|v}(w, v) = \frac{h_{wv}(w, v)}{\int_{-\infty}^{\infty} h_{wv}(x, v) dx} \quad (4)$$

for every v where the denominator in the above formula is **nonzero**. Then the conditional distribution function $F_{w|v}$ can be found explicitly by

$$F_{w|v}(d, v) = \int_{-\infty}^d h_{w|v}(w, v) dw. \quad (5)$$

For any $v \in \mathbb{R}$ the conditional quantile $q_C(p, v)$ of order p can be uniquely defined as in the unconditional case, i.e.

$$q_C(p, v) = \sup\{q \mid F_{w|v}(q, v) \leq p\}. \quad (6)$$

The following equations

$$q_C(F_{w|v}(q, v), v) = q \quad (7)$$

$$F_{w|v}(q_C(p, v), v) = p \quad (8)$$

are readily shown to hold in any point where $F_{w|v}$ is a continuous function of its first argument.

Generalization of the definition of the conditional distribution function $F_{w|v}$ for $v \in \mathbb{R}^n$ and $w \in \mathbb{R}^m$ is straightforward, as of well as the conditional quantile q_C for $v \in \mathbb{R}^n$ and $w \in \mathbb{R}$. In the most general case, where $w \in \mathbb{R}^m$, the quantiles $q(p)$ and $q_C(p, v)$ are to be defined as sets. That is omitted here.

3. Applications

Two applications of conditional distribution functions and quantiles are considered below. One is a time-optimal control problem from the area of decision theory, and the other is the task of data compression in the field of digital signal processing.

A. Time-optimal control

Let the function $l : \mathbb{R} \times \mathbb{R} \rightarrow [0, \infty)$, here after referred to as the loss function be defined by

$$l(W, w) = \begin{cases} -a(W - w) & \text{if } W - w < 0 \\ b(W - w) & \text{if } W - w \geq 0 \end{cases}. \quad (9)$$

Obviously $l(W, w) = 0$ for $W = w$, so the loss function describes the losses incurred when the estimate W does not equal the optimal value w . As seen from equation (9) losses may depend strongly on the sign of estimation error, depending on the values of the coefficients a and b . If the optimal value w is a random variable with a probability distribution P_w , the expected loss is given by the so called Bayes loss function l_B defined as follows

$$l_B(W) = \int_{\mathbb{R}} l(W, w) dP_w(w). \quad (10)$$

The value $l_B(W)$ simply constitutes the expected loss when estimating w by the value W . Any real number W_B such that

$$l_B(W_B) = \inf_{w \in \mathbb{R}} l_B(w) \quad (11)$$

is called a Bayes estimator. When the loss function l is defined by equation (9) it is readily shown that the Bayes estimator equals the quantile of order

$$p = \frac{a}{a+b}. \quad (12)$$

A practical example illustrating the relevance of quantiles as Bayes estimators is described in paper [11], where they are the solution of a time-optimal control problem. A parameter w , representing motion resistances in a mechanical system, is estimated by the value W , which appear directly in the equations of a time-optimal feedback controller. If $W > w$, overshoots occur which increases the time of reaching the target proportionally to $W - w$ with a coefficient b . In the case where $W < w$, so-called sliding trajectories appear, also prolonging the reaching period proportionally to $w - W$ with a coefficient a . The Bayes optimal estimate of the parameter

w therefore exactly constitutes a quantile of order $\frac{a}{a+b}$.

When explanatory variables v are available the Bayes estimator for the above loss function is constituted by the conditional quantile q_C . In the time-optimal control

problem, the vector v may contain disturbances possibly influencing the resistances of motion, as for example temperature or target position. Minimum expected reaching time is then obtained for $W = q_C(p, v)$.

B. Signal processing

When a finite capacity channel is used for data transmission, coding is often provided for optimal channel utilization. One example is the ADPCM [12] speech coding used for digital data transmission in mobile telephony. The objective is to transform a sequence of mutually dependent random data $\{..., u_{-1}, u_0, u_1, u_2, ...\}$ with a certain distribution P_w into a sequence $\{uc_0, uc_1, uc_2, ...\}$ of mutually independent and uniformly distributed data. The transformed data are then quantified and transmitted across the line to the receiver where they are decoded. ADPCM coding is designed on the assumption that the uncoded data exhibit the properties of linearly filtered white noise, in which case coding can be performed by the inverse linear filter removing the time correlation of the data. When the uncoded data are not correlated but still strongly dependent, linear filtering will of course work.

Assume that the sequence $\{u_i\}$ is Markov, i.e.

$$F_{u_i | u_{i-1}, u_{i-2}, \dots} = F_{u_i | u_{i-1}} = F_{w|v}, \quad (13)$$

and that the conditional distribution function $F_{w|v}$ is a continuous function of its first variable. Then the sequence $\{uc_i\}$ defined by

$$uc_i = F_{w|v}(u_i, u_{i-1}) \quad (14)$$

will be independent and uc_i is uniformly distributed on the interval $[0, 1]$ for all i . After quantization the sequence $\{ucq_i\}$ is transmitted. At the receiver side the original sequence with quantization error $\{udcq_i\}_{i \in \mathbb{N}}$ be restored by

$$udcq_i = q_C(ucq_i, udcq_{i-1}). \quad (15)$$

The coding scheme can directly be generalized to k th order Markov processes, i.e. where:

$$F_{u_i | u_{i-1}, u_{i-2}, \dots} = F_{u_i | u_{i-1}, u_{i-2}, \dots, u_{i-k}} = F_{w|v}. \quad (16)$$

The above coding scheme is based on the assumptions that the conditional distribution $F_{w|v}$ is known in both ends of the transmission line. Therefore means for transmitting this knowledge without occupying any significant amount of channel capacity is needed.

4. Neural Networks for Estimating Conditional Distributions

Feed Forward Neural Networks are most frequently

trained by applying some sort of optimization procedure like Back Propagation in order to set weights and offsets optimally w.r.t. some objective function. In most cases the objective function equals the average of some loss function on the available set of data. Thus the objective function constitutes an estimate of the expected loss function, i.e. the Bayes loss function, and the training procedure an attempt to minimize the Bayes loss function. Successful training will force the neural network output to estimate the theoretical optimum, which for mean squared error is the conditional mean, and for the loss function l defined in formula (9) equals the conditional quantile.

In this section, a neural network for solving the more general problem of estimating conditional distribution functions, and their associated quantiles of any order. In paper [13] a perceptron like structure is trained with Back Propagation to reproduce so called fractional bins representing the conditional density. Here the reasoning follows the constructive line of works [4, 5, 6] and is based on the theory of kernel estimation, which will be introduced shortly below.

A. Kernel estimation

Let $\{w_i\}$ in the following be a sequence of identically distributed random variables with a common density h_w . For any $m \in \mathbb{N} \setminus \{0\}$ and $r > 0$ the density estimate $h_w^{m,r}: \mathbb{R} \rightarrow \mathbb{R}$ can be defined by

$$h_w^{m,r}(w) = \frac{1}{mV(r)} \sum_{i=1}^m \phi\left(\frac{w - w_i}{r}\right), \quad (17)$$

where the volume function V is expressed as

$$V(r) = \int_{-\infty}^{\infty} \phi\left(\frac{w}{r}\right) dw \quad (18)$$

and the kernel function $\phi: \mathbb{R} \rightarrow \mathbb{R}$ obeys

$$\lim_{r \rightarrow 0} \frac{1}{V(r)} \int_{-\infty}^{\infty} h(w) \phi\left(\frac{w - d}{r}\right) dw = h(d) \quad (19)$$

for any bounded continuous density function h . The above estimator has been investigated in paper [14] for the case of the sequence $\{w_i\}$ being i.i.d. (independent identically distributed) random variables with a common continuous density function h_w . For $r \rightarrow 0$, and $m \cdot r \rightarrow \infty$ as $m \rightarrow \infty$, the mapping $h_w^{m,r}$ is shown to be a pointwise consistent estimator of the density h_w and its modes. By interpreting the kernel function ϕ as the nonlinear function of a neuron, and the sequence $\{w_i\}$ as a set of observations serving as training data, it has been demonstrated in work [4] how this estimator exhibits properties equivalent to neural networks. From a computational point of view it possesses a massively

parallel structure, which allows for high speed implementation on dedicated hardware; functionally, it is capable of learning general probabilistic information from measured data. It should be pointed out, however, that the number of neurons in the network defined from formula (17) equals the number of data in the training set, and that learning more or less takes place by memorizing data. In that respect the network provides no data compression.

In papers [5, 6] the estimator (17) was transformed to compute conditional expectation functions and a structure equivalent to the kernel smoother described in work [15] was obtained. Here this transformation is directed towards estimators of conditional distribution functions.

In the multivariable case the training data is a finite sequence of the form $\{(w_i, v_i)\}$, where v_i denotes an observation of some observable explanatory variable. In that case the multivariable density estimate $h_{wv}^{m,r}$ can be given as

$$h_{wv}^{m,r}(w, v) = \frac{1}{mV(r)} \sum_{i=1}^m \phi\left(\frac{w - w_i}{r}\right) \cdot \phi\left(\frac{v - v_i}{r}\right). \quad (20)$$

A conditional distribution estimate $F_{w|v}^{m,r}$ can be obtained by subjecting the function $h_{wv}^{m,r}$ to a transformation analogous to the one defined by equations (4) and (5), i.e.

$$F_{w|v}^{m,r}(d, v) = \frac{\int_{-\infty}^d h_{wv}^{m,r}(w, v) dw}{\int_{-\infty}^{\infty} h_{wv}^{m,r}(w, v) dw}, \quad (21)$$

which leads to the following closed form expression

$$F_{w|v}^{m,r}(d, v) = \frac{\sum_{i=1}^m S\left(\frac{d - w_i}{r}\right) \cdot \phi\left(\frac{v - v_i}{r}\right)}{\sum_{i=1}^m \phi\left(\frac{v - v_i}{r}\right)}, \quad (22)$$

where S denotes the antiderivative of the function ϕ :

$$S(d) = \int_{-\infty}^d \phi(w) dw. \quad (23)$$

A scaled Gaussian density function may be proposed as a candidate for the function ϕ , that is

$$\phi(d) = \exp(-|d|^2), \quad (24)$$

where $|\cdot|$ denotes the Euclidean metric. This function exhibits all properties required except that its antiderivative is not computable in a closed form expression. Therefore the function S can be chosen not according to equation (23) but as a function exhibiting equivalent properties and which is computable in a closed form expression. The well known sigmoid function then constitutes a natural choice, i.e.

$$S(d) = \frac{1}{1 + \exp(-d)}. \quad (25)$$

The above elaboration has been based on kernel estimation of a joint density function h_{wv} and leads to an estimate $F_{w|v}^{m,r}$ of the conditional distribution $F_{w|v}$; and serves here merely as motivation to formula (22). In fact, from definitions (24) and (25) it can be shown by fairly standard means, on only a very mild set of assumptions, that a slightly modified version of the mapping $F_{w|v}^{m,r}$ consistently estimates the distribution $F_{w|v}$, as stated precisely in the following theorem, which is proved in paper [16].

Theorem

Let P_{wv} be a probability measure on the space \mathbb{R}^2 with an associated distribution function F_{wv} , and define the measure P_v on the real space by

$$P_v(A) = P_{wv}(\mathbb{R} \times A). \quad (26)$$

Assume a discrete time random process $z = (w, v) : \Omega \times \mathbb{N} \rightarrow \mathbb{R}^2$ to be such that empirical distributions converge to the function F_{wv} at every point of its continuity, i.e. w.p.1

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m U(z - z_i(\omega)) = F_{wv}(z) = \int U(z - (w, y)) dP_{wv}(w, y) \quad (27)$$

for every continuity point z of the mapping F_{wv} , where the function $U : \mathbb{R}^2 \rightarrow [0, 1]$ is given as

$$U(x_1, x_2) = \begin{cases} 1 & \text{if } x_1 \geq 0 \text{ and } x_2 \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (28)$$

Also let the conditional distribution function $F_{w|v} : \mathbb{R}^2 \rightarrow [0, 1]$ fulfill the following smoothness condition

$$\lim_{v \rightarrow \tilde{v}} F_{w|v}(\tilde{d}, v) = F_{w|v}(\tilde{d}, \tilde{v}) \quad (29)$$

at any point (\tilde{d}, \tilde{v}) , where $F_{w|v}$ is continuous w.r.t. its first argument. Then, for the estimator $F_{w|v}^{r,m}$ defined by

$$F_{w|v}^{r,m}(d, v) = \frac{\sum_{i=1}^m S\left(\frac{d - w_i + \sqrt{r}}{r}\right) \cdot \phi\left(\frac{v - v_i}{r}\right)}{\sum_{i=1}^m \phi\left(\frac{v - v_i}{r}\right)} \quad (30)$$

the following is true for any number v in the support of the measure P_v :

$$\lim_{r \rightarrow 0} \left[\lim_{m \rightarrow \infty} F_{w|v}^{r,m}(d, v) \right] = F_{w|v}(d, v) \quad \text{w.p.1.} \quad (31)$$

In above Theorem only a very general ergodic property has to be fulfilled, requiring empirical measures of the data sequence to converge to the limit measure P_{wv} , which not even locally is assumed to possess a density function. The conditional distribution function $F_{w|v}$ needs to vary smoothly w.r. t. the explanatory variable v in the sense stated in equation (29), which allows for almost any degree of discontinuity w.r.t. the explained variable w . The estimator $F_{w|v}^{m,r}$ is redefined in formula (30) from its original definition in equation (22). The term \sqrt{r} now appears in the argument of the function S allowing for convergence to points of discontinuity w.r.t. the explained variable w . It should be noted that the Theorem is easily generalized to arbitrary dimensions only by redefining the function S to an arbitrary dimension D , i.e.

$$S(d_1, d_2, \dots, d_D) = \prod_{i=1}^D S(d_i). \quad (32)$$

For $D = 1$ the conditional quantile estimate $q_C^{m,r}$ is defined uniquely by

$$F_{w|v}^{m,r}(q_C^{m,r}(p, v), v). \quad (33)$$

B. Data compression

The neural network defined by equation (30) maps its training data directly onto its network parameters implying a number of neurons equal to the number of training data m . At least in the signal processing application discussed earlier some sort of efficient data compression is needed for the network to serve its purpose. Data compression is generally introduced by replacing the estimator $F_{w|v}^{m,r}$ with its compressed modification $F_{com}^{n,r}$

$$F_{com}^{n,r}(d, v) = \frac{\sum_{i=1}^n S\left(\frac{d - x_i^w + \sqrt{r}}{r}\right) \cdot \phi\left(\frac{v - x_i^v}{r}\right)}{\sum_{i=1}^n \phi\left(\frac{v - x_i^v}{r}\right)}. \quad (34)$$

In this equation n denotes the number of neurons which is considered to be a design parameter restricted by $n \ll m$ in order to ensure a sufficient level of compression. The parameters $x_i = (x_i^w, x_i^v)$, are viewed as adjustable weights and offsets, subject to some training procedure projecting the statistical information of the training data to the network parameters. Two different procedures for setting the parameters x_i are discussed below.

Algorithm 1:

- The parameters x_i are initially drawn randomly according to the joint density estimate $h_{wv}^{m,r}$ defined in

equation (20). For every $i \in 1, \dots, n$, j is selected from the numbers $1, \dots, m$ with equal probability for all values. The parameter x_i is then initially set by

$$x_i = w_j + e, \quad (35)$$

where e is independent and Gaussian distributed with a variance $\frac{A/k}{2}$.

- All parameters x_i are then submitted to an LVQ correctionscheme:
 - repeat
 - * for $j = 1, \dots, m$
 - ♦ if $|x_i - w_j| < |x_k - w_j|$ then
 - $x_i = x_i + \eta \cdot (w_j - x_i)$
 - until all parameters have stabilized.

The initial random setting according to the estimated density produces a rather qualified starting guess for the subsequent self organizing, which in turn diminishes the randomness from the initial settings to a level fair to the number of available training data n . That is the statistical uncertainty of the compressed modification $F_{com}^{n,r}(d, v)$

will be comparable to that of the estimator $F_{w|v}^{m,r}(d, v)$.

For the signal processing application discussed earlier only the initial settings might be transmitted whereas the subsequent LVQ correction were to be done independently at transmitter and receiver ends respectively. Such an approach closely resembles the ADPCM scheme mentioned earlier and even generalizes certain elements of that method.

Another approach to setting the parameters $\{x_i\}$ utilizes the estimator $F_{w|v}^{m,r}$ as well as an analogous estimator $F_v^{m,r}$, and is based on the following reasoning. Let the function $G : [0, 1]^2 \rightarrow \mathbb{R}^2$ be defined by

$$G(x^1, x^2) = (q(x^1), q_C(x^2, q(x^1))), \quad (36)$$

where q is a quantile associated to the distribution F_v and q_C is a conditional quantile associated to the conditional distribution $F_{w|v}$. If the random variables x_1 and x_2 are independent and uniformly distributed on the interval $[0, 1]$, then the random variable $G(x_1, x_2)$ is distributed according to the function F_{wv} on the space \mathbb{R}^2 . Also if $\{A_k\}$ is a sequence of finite subsets of the set $[0, 1]^2$ and the empirical distributions of the points in $\{A_k\}$ converges weakly to the uniform distribution on the set $[0, 1]^2$, then the empirical distributions of the random variables $\{G(A_k)\}$ converge weakly to the distribution

function F_{wv} . For example $\{A_k\}$ might be crossing points on rectangular grids with gridsizes converging to zero. These properties are used in Algorithm 2 to make network parameters imitate the empirical distributions of the data.

Algorithm 2:

- Select $\{x_i^1\}$ $i = 1, \dots, n_1$ equidistantly on the interval $[0,1]$.
- Let $y_i = q^{m,r}(x_i^1)$ for $i = 1, \dots, n_1$.
- Select $\{x_j^2\}$ $j = 1, \dots, n_2$ equidistantly on the interval $[0,1]$.
- for $i = 1 : n_1$
 - for $j = 1 : n_2$
 - * $k = n_2 \cdot i + j$
 - * $x_k^v = y_i$
 - * $x_k^w = q_C^{m,r}(x_j^2, y_i)$

The parameter setting scheme above can be generalized to arbitrary dimensions but it is not recommended for high dimensionality as the computational effort tends to grow exponentially with dimension. The advantage of this second approach for parameter setting is that constitutes a deterministic mapping from the available training data onto the network parameters, which makes it feasible for analysis. Convergence properties for the compressed estimator $F_{com}^{n,r}$ could be stated along the same line as for the estimator $F_{wv}^{m,r}$. The possibility for such analysis is only mentioned at this point, whereas the analysis itself is not presented in this paper.

5. Conclusion

A neural network for estimating conditional distributions and their associated quantiles has been constructed in the present paper. Although the network is designed on the basis of kernel estimation of joint probability density functions, theory has been presented showing the network to be valid in more general settings, where only a smoothness condition w.r.t. the dependence on the explanatory variable, as well as a very general ergodic property of the training data have to be fulfilled.

The problem of estimating conditional quantiles has been related to Bayes estimation in the case of a special asymmetric loss function feasible for application within a variety of areas in engineering, as well as science and economics. An example from the area of time-optimal control is briefly discussed.

An application of the presented neural network within digital signal processing has been suggested. A scheme for encoding and decoding a sequence of data for optimal

channel utilization is presented, along with two algorithms for training the network parameters before and during data transmission.

References

- [1] D.E.Rumelhart and J.McClelland. *Parallel Distributed Processing*, MIT Press, 1986.
- [2] H.White. Connectionist Nonparametric Regression: Multi Layer Feedforward Network can Learn Arbitrary Mappings, *Neural Networks*, 3: 535-549, 1990.
- [3] D.W.Ruck, S.K.Rogers, M.Kabrisky, M.E.Oxley, and B. W.Suter. The Multilayer Perceptron as Approximation to a Bayes Optimal Discriminant Function, *IEEE Transactions on Neural Networks*, 1: 296-298, 1990.
- [4] D.F.Specht. Probabilistic Neural Networks for Classification, Mapping, or Associative Memory, *Proc. IEEE International Conference on Neural Networks*, 1: 525-533, 1988.
- [5] D.F.Specht. A General Regression Neural Network, *IEEE Transactions on Neural Networks*, 2: 568-576, 1991.
- [6] H.Schiöler and U.Hartmann. Mapping Neural Network Derived from the Parzen Window Estimator, *Neural Networks*, 5: 903-909, 1992.
- [7] J.Moody and C.Darken. Fast Learning in Networks of Locally Tuned Processing Units, *Neural Computation*, 1: 28 1-294, 1989.
- [8] R.H.Nielsen. Counter Propagation Networks, *Proc. First IEEE International Conference on Neural Networks*, 2: 19-33, 1987.
- [9] M.Fisz. *Probability Theory and Mathematical Statistics*, Wiley, 1963.
- [10] P.Billingsley. *Probability and Measure*, Wiley, 1979.
- [11] P.Kulczycki. Time-Optimal Stochastic Positional Control, *Proc. IFAC 12th World Congress*, 7: 443-448, 1993.
- [12] S.Haykin. *Communications Systems*, Wiley, 1983.
- [13] A.S. Weigend and A.N. Srivastava. Predicting Conditional Probability Distributions: A connectionist approach, *International Journal of Neural Systems*, 6, 1995.
- [14] E.Parzen. On Estimation of a Probability Density Function and Mode, *Annals of Mathematical Statistics*, 33: 1065-1076, 1962.
- [15] T. J.Hastie and R. J.Tibshirani. *Generalized Additive Models*, Chapman and Hall, 1990.
- [16] H.Schiöler and P.Kulczycki. Neural Networks for Estimating Condition Distributions, *IEEE Transactions on Neural Networks*, 8, 1997, submitted for publication.