Delft, June 13 2006

Delft University of Technology
Faculty of Electrical Engineering,
Mathematics and Computer Science

EXAMINATION NEURAL NETWORKS (IN 4015)

Wednesday, June 21 2006, 14.00 – 17.00

Answer exactly what is questioned and try to formulate as precisely as possible !! Provide your answers in Dutch or English.

Mention on every sheet of paper your name and study number !!

I. Data fitting with linear models

1. In the case of a supervised adaptive system we have to minimize an error criterion. Frequently, the so-called MSE (= Mean Squared Error) is used as error criterion function, which normally is indicated by J(w) where w is the weight vector. The w-vector for which J takes a minimum, say w*, can be found by taking: grad J = 0. This leads to the so-called normal equations which in vector notation can be formulated as

$$p = R\, w^*$$

where p is a cross-correlation vector and R is an autocorrelation matrix. Please, indicate what exactly p and R are, in other words specify their elements.

2. Instead of the analytical method (see the above question 1) to minimize the MSE error criterion, in practice search methods are used. A particularly elegant method is the so-called LMS (Least Mean Square) method to search for the minimum. Give a mathematical expression for the LMS search method (for the multi-dimensional case); derive this LMS search method from the well-known steepest descent method (gradient method).

3. Consider the steepest descent search method. Theoretical results for the largest stepsize η and the "time constant of weight adaptation" τ are:

$$\eta < 2 / \lambda_{max} \quad \text{and} \quad \tau = 1 / \eta\lambda_{min}$$

What does λ mean in these equations? Explain why in the inequality for η the *maximum* value of λ is taken while in the equation for τ the *minimum* value of λ is taken.

II. Pattern recognition (classification)

4. A vector x is said to belong to class c_i if

$$p(x \mid c_i) \cdot P(c_i) > p(x \mid c_j) \cdot P(c_j) \quad \text{for all } j \neq i$$

Explain this (taking into account Bayes' rule). How do you determine in practice the likelihood $p(x \mid c_j)$ and prior probability $P(c_j)$?

5. Let's assume $p(x \mid c_i)$ in the foregoing question 4 concerns a Gaussian distribution and x is 2-dimensional. The Gaussian then is characterized by a 2-dimensional mean and a 2 x 2 covariance matrix. What exactly is put in the rows and columns of this covariance matrix? What is the special feature of the matrix for a sufficiently large number of measurements?

6. Explain what is meant with "discriminant functions" and how with these discriminant functions a general parametric classifier for c classes can be constructed (give also a schematic figure of the classifier).

7. What is the difference between "parametric classifiers" and "nonparametric classifiers"? Neural networks belong to an intermediate form of these two different kinds of classifiers. Why?

III. Multilayer perceptrons (MLPs)

8. Consider a single McCulloch-Pitts neuron (or processing element) M-P PE with three inputs (x,y,z) with associated weights w_1, w_2 and w_3 as well as the bias weight b. The M-P PE is used for classification in two (2) classes in the 3-dimensional input space (x,y,z). The decision surface in this input space is characterized by the function z = f(x,y). Give the mathematical expression of this function f(x,y).

9. The so-called *modified* M-P PE has a non-linear activation function. Examples of frequently used activation functions are the logistic function and the tanh-function. Give the mathematical expression of the logistic function, its first derivative, and the first derivative of the tanh-function. Take into account a steepness parameter α in all formulas!!

10. Compare the learning rule for a M-P PE (Rosenblatt 1958) with the so-called Delta rule for a *modified* M-P PE (Note: there are two major differences !!).

11. What is meant with a one-layer perceptron? Explain on the basis of a one-layer perceptron consisting of n M-P PEs that its decision regions in the n-dimensional input space are always convex regions.

12. For training an MLP the weight adaptation can be generally expressed as follows:

$$\Delta w_{ij} = \eta \, \delta_i \, y_j$$

where w_{ij} is the weight associated to the j-th input of the considered i-th neuron and y_j is the output of a j-th neuron linked to the considered i-th neuron (so actually the j-th input signal of the considered i-th neuron).
What is δ_i **A)** for the case that the considered i-th neuron has a linear activation function and is at the output layer, **B)** for the case that the considered i-th neuron has a non-linear activation function and is at the output layer, and **C)** for the case that the considered i-th neuron has a non-linear activation function and is at a hidden layer ?

13. The backpropagation algorithm can be considered as a flow of activations in forward direction through the neural network and a flow of (injected) errors in backward direction through the so-called dual network. Show that the splitting nodes in the original network change into summing junctions in the dual network, and vice versa (please, use a figure for explanation).

IV. Designing and training MLPs

14. The choice of the variance of the initial values of the weights of an MLP, before starting the training process, should be proportional to the inverse of the so-called "fan-in" (i.e. number of inputs) of a neuron. Explain why.

15. Give the weight-update rule for MLPs in the case of "momentum learning"? Which are the advantages to be expected with momentum learning?

16. Frequently, the criterion to stop training of an MLP is based on the generalisation to be achieved. Show in a figure and discuss that it is not good to continue training too long (in order to avoid so-called overtraining).

17. The classifying performance of an MLP-classifier can be made clear in a so-called "confusion matrix". Discuss this matrix for the case of classification in four (4) classes. How can one deduce from the matrix the accuracy of the classifier?

18. Instead of the Mean Squared Error (MSE) criterion sometimes the so-called Cross-Entropy criterion is used. It appears that with this Cross-Entropy criterion the normal back-propagation procedure can be applied, however with two small modifications. Which are these modifications?

19. Advanced search methods for finding a minimum of the error criterion J(w) are frequently based on Newton's method. Give a mathematical expression of Newton's method; which are the most essential features of the method?

20. Pseudo-Newton's methods are also used to replace the pure Newton method in practice. The idea is to come up with computationally simple but reasonable approximations of the Hessian. Discuss this briefly and give an example of such approximation in mathematical terms.

V. Competitive and Kohonen Networks

21. What is the role of the so-called "winner-take-all network" in the hard competition version of competitive learning (present in your answer also the architecture of this network)?

22. The criterion for competition in competitive learning frequently is the similarity between the input vector and weight vector per neuron. This similarity can be estimated on the basis of the Euclidian distance or the angle between the vectors concerned. Discuss both methods with use of mathematical formulas.

23. How can be avoided that during competitive learning always the same neuron(s) win(s) the competition? Describe the procedue in mathematical terms.

24. Describe the so-called Kohonen training with use of the required mathematical formulas.

VI. Neural Control

25. A so-called "neural identifier" is part of many neural control systems. What is a neural identifier ?

26. Explain the necessity of so-called TDLs (Tapped Delay Lines) when using a backpropagation network as neural identifier.

27. What is Forward-Modelled Inverse Control? Give and discuss the control configuration and explain the role of the neural identifier herein.

==========