

AD-A174 388

A PATTERN RECOGNITION APPROACH TO UNDERSTANDING THE

1/1

MULTI-LAYER PERCEPTOR. (U) ROYAL SIGNALS AND RADAR

ESTABLISHMENT, MALVERN (ENGLAND) J D LONGSTAFF ET AL.

UNCLASSIFIED

JUL 86 RSRE-MEMO-3936 DRIC-WR-100622

F/G 12/1

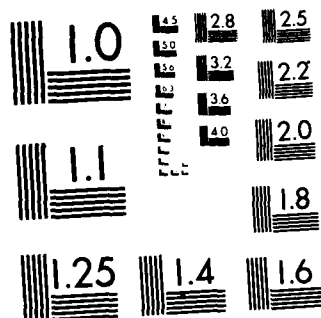
NL

END

DATE

FILED

1 87



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963-A

UNLIMITED



**RSRE
MEMORANDUM No. 3936**

AD-A174 308

**ROYAL SIGNALS & RADAR
ESTABLISHMENT**

A PATTERN RECOGNITION APPROACH TO
UNDERSTANDING THE MULTI-LAYER PERCEPTRON

Authors: I D Longstaff and J F Cross

**PROCUREMENT EXECUTIVE,
MINISTRY OF DEFENCE,
RSRE MALVERN,
WORCS**

UNLIMITED

ROYAL SIGNALS AND RADAR ESTABLISHMENT

Memorandum 3936

TITLE: A PATTERN RECOGNITION APPROACH TO UNDERSTANDING
THE MULTI-LAYER PERCEPTRON

AUTHORS: I D Longstaff and J F Cross

DATE: July 1986

ABSTRACT

This memorandum is concerned with the operation of a class of multi-layer associative networks commonly known as the multi-layer perceptron (MLP), Rumelhart network or back-propagation network. We describe the operation of the MLP as a pattern recognition device in terms of a feature-space representation. This allows an understanding of how structure in the training data is represented internally in the machine.

Index Terms - Perceptron, neural networks, associative networks, pattern recognition, feature-space, error back-propagation.

Copyright
C
Controller HMSO London
1986

Accession For	
NTIS CRA&I	☒
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and/or Special
A-1	

RSRE MEMORANDUM 3936

A PATTERN RECOGNITION APPROACH TO UNDERSTANDING
THE MULTI-LEVEL PERCEPTRON

I.D. Longstaff and J.F. Cross

CONTENTS

I INTRODUCTION

II FEATURE-SPACE REPRESENTATION

A. The Two-Class Problem

- 1) Perceptron with no hidden layers.
- 2) Perceptron with one hidden layer.
- 3) Perceptron with two hidden layers.

B. The Multi-Class Problem

III TRAINING

IV SUMMARY

V REFERENCES

I INTRODUCTION

The perceptron has long been known as a device which can map input patterns from different classes to output patterns representing the class identification. Minsky and Papert [1] showed that an associative network with no hidden layers, for which there was a training algorithm, could not perform all possible mappings and some patterns, such as the exclusive-or, could not be recognised. On the other hand a multi-layer machine could perform all mappings, but there was no training algorithm.

Recently interest in the multi-layer perceptron has revived with the introduction of a deterministic training algorithm by Rumelhart et al [2] which allows convergence to a solution. Our experience with using these networks for pattern recognition tasks showed that a better understanding of their operation is required in order to have any confidence that the system operates efficiently on data sets with a large number of variates or classes. Some guidelines are required on the number of hidden layers and nodes needed for a particular problem. Also we found that training was not effective on problems with large numbers of classes and a solution to this problem was required.

II FEATURE-SPACE REPRESENTATION

We use here the notation commonly employed in statistical pattern recognition whereby the set of features or variates used to classify a pattern are represented as the axes of an N-dimensional Cartesian coordinate system or 'feature-space'; see

for instance Duda and Hart [3].

Individual patterns (exemplars) are then represented as individual points in feature-space. Decision boundaries, positioned using the training data, allow regions in feature-space to be associated with a particular class. A new data point can be classified according to the region within which it falls. Associative networks are of interest because of their ability to find multiple correlations between the variates of the input vectors and to make generalisations from a limited number of observations so that previously unseen inputs can be associated with earlier inputs. These abilities can be explained in terms of the feature-space representation. Multiple and high order correlations between features of a particular pattern class are embodied in the shape of the region (or regions) of feature-space representing the pattern class. The process of making generalisations from a few observations is identical to associating new points with the region around the initial training observation. This usually entails forming decision boundaries around the cluster (or clusters) of training points representing each class of pattern.

For a multi-class problem it is important to consider the type of classifier required as this can influence the training procedure. A common requirement is for a maximum-likelihood classifier where each region of feature-space is labelled with only one class. An alternative is to label each region with the combination of all plausible classes, for instance as might be

required as an input to a subsequent knowledge based system.

For the time being we only consider the two-class problem and then show how this can be extended to either of the multi-class problems.

We give now a step-by-step interpretation of the operation of the MLP in terms of the feature-space representation starting at the input layer.

A. The Two-Class Problem

1) Perceptron with no hidden layers: Fig. (1) shows a single node perceptron above the input level. The components x_i of the input vector $\underline{x}$ are multiplied by the elements w_i of a weight vector $\underline{w}$ and summed with a bias term w_b . Rumelhart [2] has shown that if the node output is a non-linear function (with certain properties) of this sum then multi-layer networks, formed by these nodes, can be trained to recognise patterns. The usual non-linear function is:

$$s = 1/[1+\exp(-E)]$$

where s is the output state

$$\text{and } E = \sum_i w_i x_i + w_b$$

A hyperplane decision boundary in feature-space ($\underline{x}$) is formed from a threshold on the output. This threshold is normally taken as 0.5.

Since $s=0.5$ when $E=0$ the hyperplane is given by $\sum_i w_i x_i + w_0 = 0$

The non-linear function causes the output of the node to approach zero or one except if the input vector is near the decision boundary. A set of nodes at the first layer defines a set of planes in ($\underline{x}$) feature-space, as in Fig. (2). The input data is mapped onto the corners of a unit hypercube in $\underline{s}$ space if the data is well away from the decision boundary in $\underline{x}$ space. If the input data is near a decision boundary then it will appear within an edge, face, or volume of the cube; depending on how many hyperplanes the point is near. For the time being it is convenient to use the simplification that all points appear at the corners of the hypercube and each region in $\underline{x}$ space is therefore defined by a unique binary code.

2) Perceptron with one hidden layer: A single node above the first hidden layer forms a decision plane in $\underline{s}$ space. This can slice any corner, edge, face, etc. off the cube. There is little point in slicing edges, faces or higher dimensional sets of this type because this would mean the classifier is indifferent to the features (in $\underline{s}$) comprising the edge or face etc. and so these features in $\underline{s}$ can be disregarded.

A dissection of particular interest is that which slices off a selected corner. This is equivalent to recognising just one binary code (at a single node above the first hidden layer) as class A, all other codes being of class not-A. In this form the classifier defines a single convex region in $\underline{x}$ space, perhaps unbounded. Other shapes with concave parts to the surface, or

disconnected regions cannot always be represented; although the MLP will often make a good attempt.

Thus a bound on the performance of the MLP with one hidden layer and one output node is that a single convex decision surface can be guaranteed. The rounding of corners shown on the decision boundary in Fig. (2) comes about because the hidden node outputs are summed at the output and if two or more are near the transition point a smaller contribution is required from each to reach the switching threshold.

The number of nodes in the first hidden layer determines the accuracy with which a required decision boundary can be represented. For instance, at least $N+1$ nodes are required for a closed volume in N -space. This would give a pyramid shape (with rounded corners) whereas $2N$ nodes would allow box like shapes (with rounded corners).

If the training data set size is small, care must be taken not to specify too many nodes at the first hidden layer else the decision boundaries will follow insignificant detail of the sample distribution. Foley [4] discusses the training sample size required for pattern recognition problems and indicates that each class should have a sample size at least three times the number of features (variates) for multivariate normal distributions. Clearly this number would have to be multiplied by the number of modes if a pattern distribution was multimodal.

3) Perceptron with two hidden layers: By specifying a number of nodes above the first hidden layer we have a structure where each output can turn on in response to an input which falls inside the region defined by that output node. If the outputs of the second layer of nodes are combined into a single output node, as in Fig. (3), we can form the union of the separate regions as for instance may be required for a pattern with a bimodal distribution.

It should be noted that if the distribution of training points in ($\underline{x}$) space is bimodal such that two enclosed regions are required then the numbers of nodes at the first hidden layer would need to be $2(N+1)$ or more.

Since any realisable shape, disjoint or concave, can be approximated from the union of (possibly overlapping) convex regions we have a fundamental upper limit to the complexity of the MLP for the two class problem. This limit is that a MLP with two hidden layers can solve any pattern recognition problem.

B. The Multi-Class Problem

We can now replicate the structure in Fig. (3) to test for the presence or absence of features representing other classes, with each output node representing each class ie A or not-A, B or not-B, C or not-C, etc. With this type of classifier it is possible for more than one of the outputs to be turned on at once, indicating that the input vector has the attributes of more than one class. This may or may not be desirable in a particular implementation of a pattern classification system, but the extent

to which this occurs is determined by the way the classes of not-A, not-B, etc, are defined at the training stage.

If overlap regions are to be avoided in a particular application of the MLP, these can be minimised by using a mixture of classes B, C, etc, as the not-A class during training of the class A classifier. The proportion A, B, etc in the mixture should be in the ratio of their a-priori probabilities, if these were available.

If overlap regions are required, for instance to give an indication of possible confusion, then a suitable prior distribution for the class not-A would be a uniform distribution in x space. In this way a decision about possible membership of one class is not influenced by any decisions regarding the other classes.

We see therefore that the network shown in Fig. (4) represents an MLP structure which is guaranteed to represent all multi-class problems, however complex.

III TRAINING

Little has been said about training, but since we now have an understanding of one internal representation which can perform all pattern recognition tasks we are able to assist the learning process. By decomposing from a multi-class into many two-class problems we are able to consider each two-class problem in turn; this may add to the number of nodes required for the complet

system but it does force an understandable internal representation and limit the extent of the training problem to the pairwise case and allow training with one pair of classes at a time.

The approach used here throws some light on situations which cause the training algorithm to stick in local minima when searching for the global minimum error. For instance consider the problem of forming a decision surface around two separate clusters which are to have the same class designation as in Fig.(5). If this disjoint characteristic of the data was not known a priori the MLP would have to discover it. At the start of the training process a random set of weights is selected; this places a set of random decision hyperplanes in the input feature-space. If, by chance, some of the surfaces intersect the space between the two clusters the gap will be discovered. If on the other hand the gap is not intersected the decision surfaces may only close in and surround the joint set. To discover the gap at least two of the boundaries would have to move in a direction which increased the error rate, which does not happen with Rumelhart's deterministic procedure. By repeating the training process with new random weights each time the probability of discovering the optimum solution can be increased. The randomisation suggested here to overcome the local minima problem is reminiscent of the randomisation which occurs when the Metropolis algorithm is used to train stochastic associative networks such as the Boltzmann machine [5]. An alternative to repeated random starts would be to use a large number of nodes,

so increasing the chance of a decision boundary being placed in the gap at the start of training. Also, any nodes which become isolated (weights approaching zero) should be re-started at a new random position.

IV SUMMARY

The operation of the multi-level perceptron (MLP) has been described in terms of a feature-space representation which allows the structure of the data to be related to structure of the network.

It is shown that a pattern recognition task with any number of classes, with analogue features and with any degree of complexity can be solved by the MLP using only two hidden layers. It is made clear that the number of nodes at each layer relates to the complexity of the problem in terms of both the number of classes and the detail required to form decision boundaries.

The insight given by these observations indicates that a structured training programme may overcome the difficulty often encountered with these networks: namely the failure to find a solution when the training algorithm is applied with large numbers of classes. Examples are given where the system can fail to learn, even with a simple problem, and a method for overcoming this is suggested.

V REFERENCES

[1] M. L. Minsky and S. Papert, Perceptrons. Cambridge, MA: MIT Press, 1969.

[2] D. E. Rumelhart, G. E. Hinton and R. J. Williams. "Learning internal representations by error propagation," ICS Report 8506, University of California, Sept. 1985.

[3] R. O. Duda and P. E. Hart, Pattern Classification and Scene Analysis. New York: Wiley, 1973. pp. 3-4.

[4] D. H. Foley, "Consideration of sample and feature set size," IEEE Trans. Inform. Theory, vol. IT-18, No.5, pp. 618-626, Sept. 1972.

[5] D. H. Ackley, G. E. Hinton and T. J. Sejnoeski, "A learning algorithm for Boltzmann machines," Cog. Sci., vol 9, pp. 147-169, 1985.

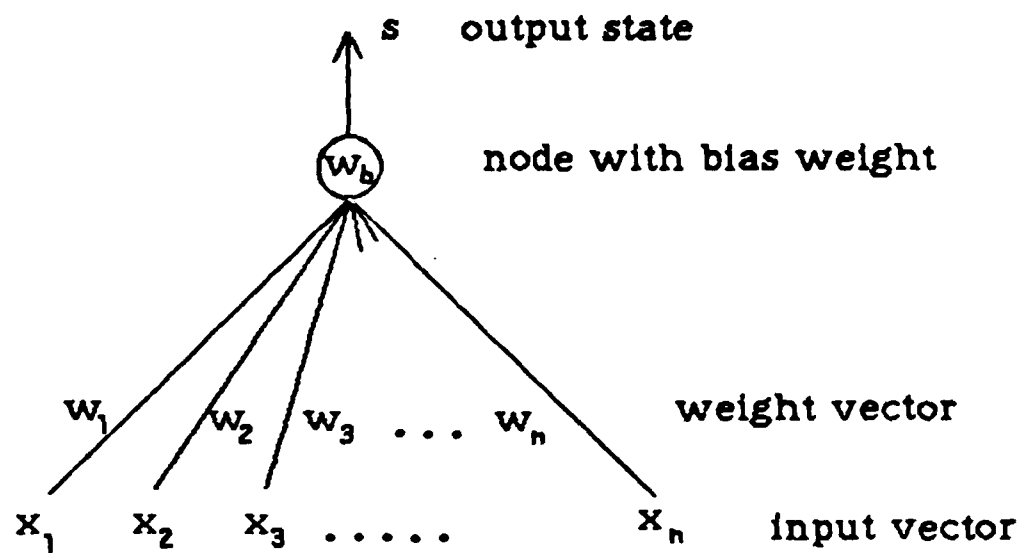


Figure 1 Single Node Perceptron.

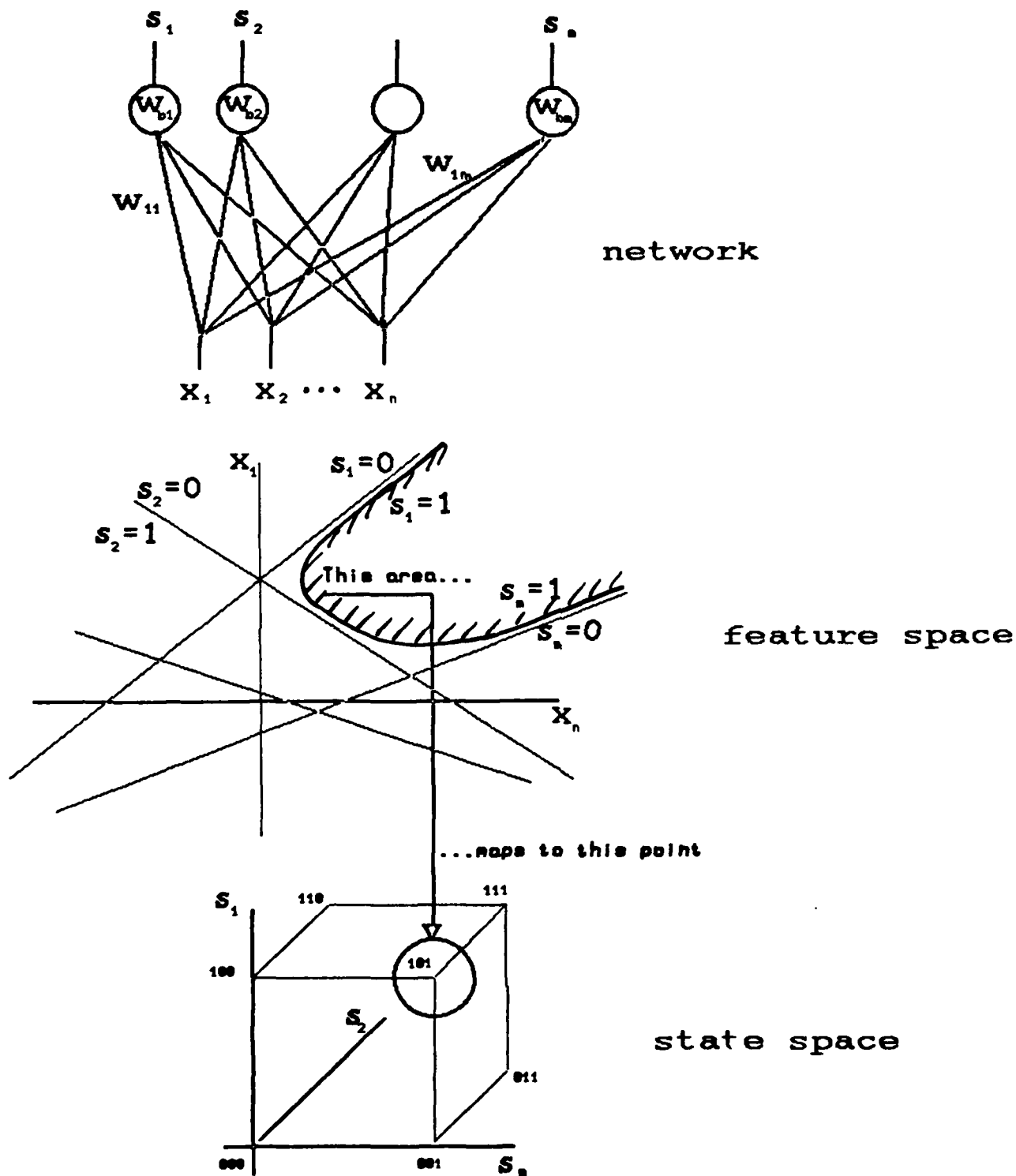


Figure 2 State Space With a Number of Nodes Above the Input Layer

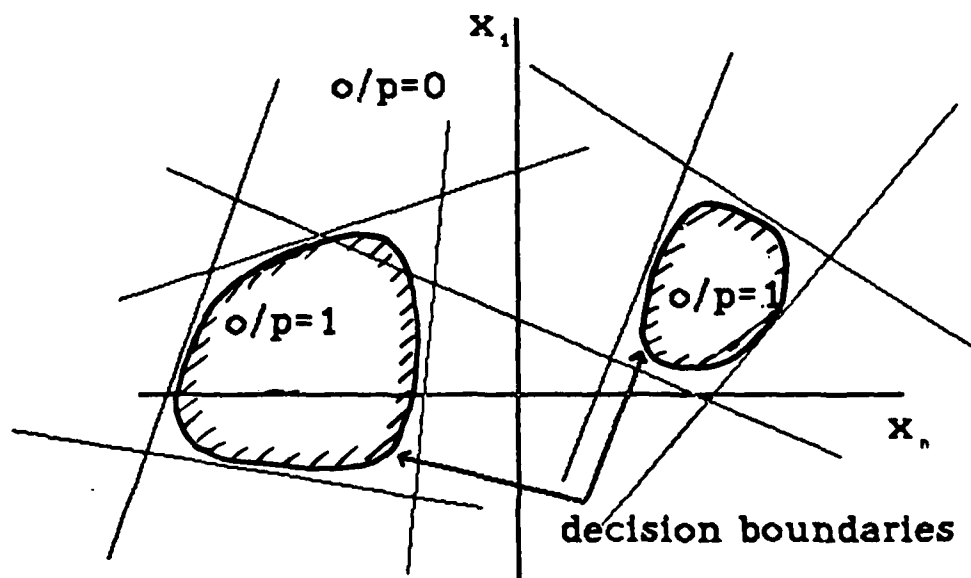
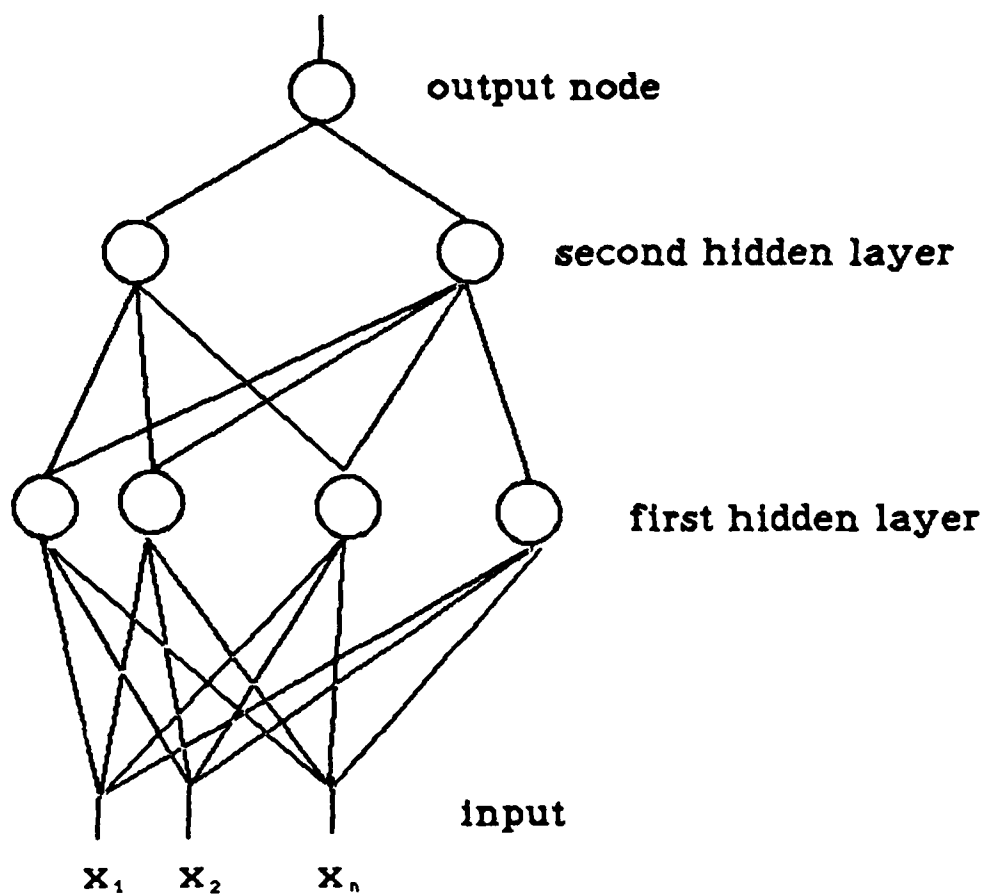


Figure 3 Perceptron With Two Hidden Layers.

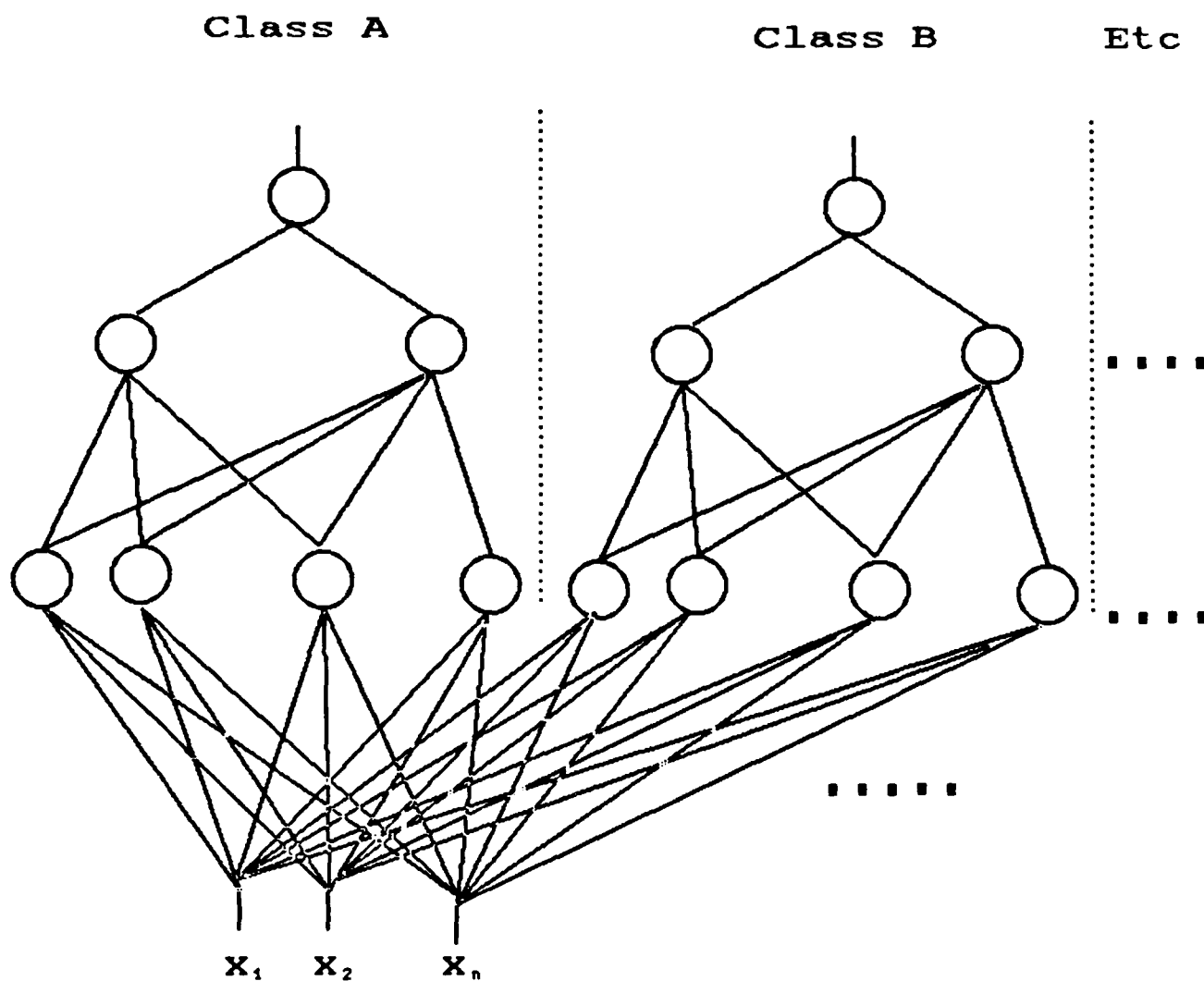


Figure 4 Multi-class perceptron.

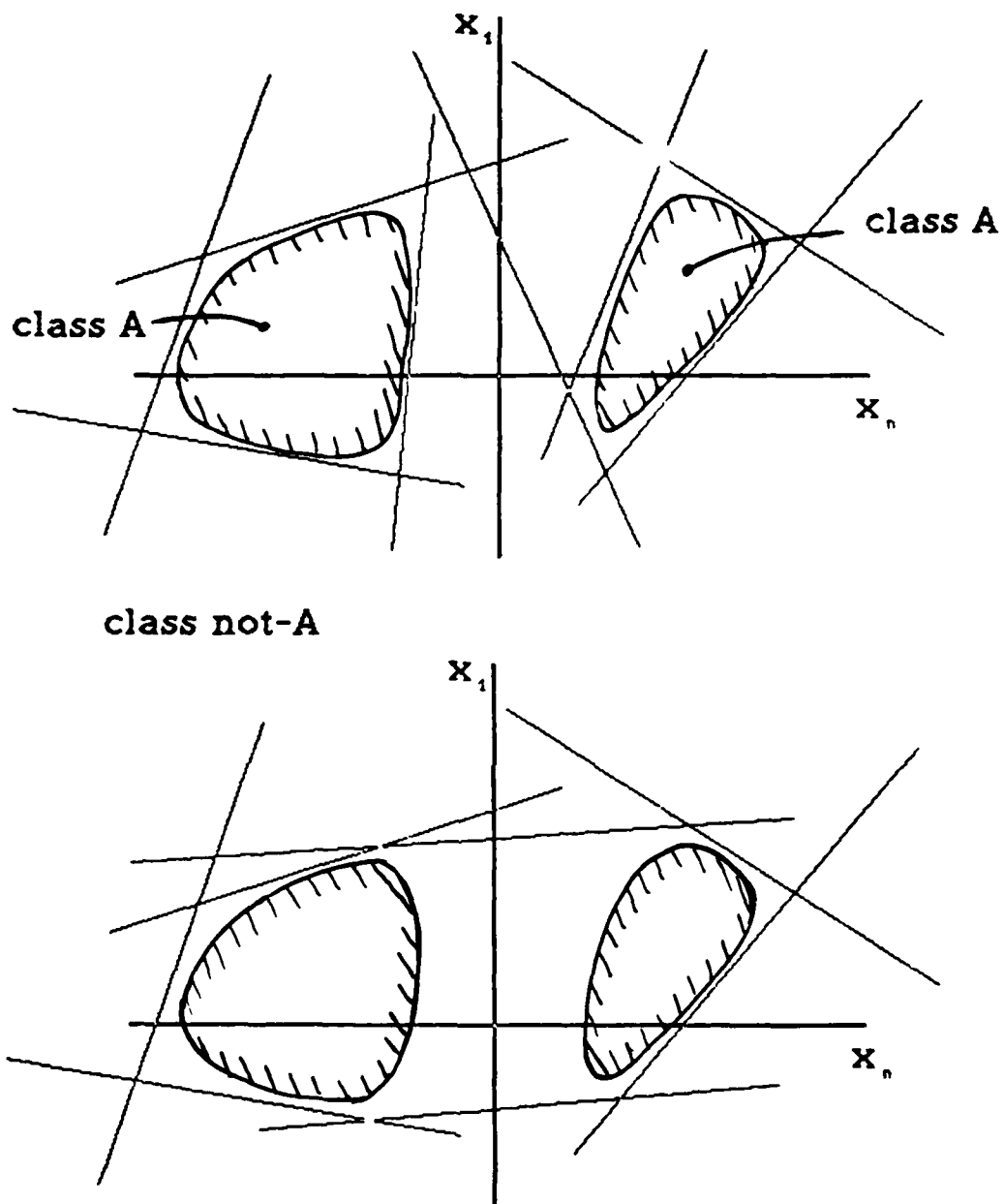


Figure 5 A Bimodal Distribution Problem

a) Gap Found

b) Gap Not Found

DOCUMENT CONTROL SHEET

Overall security classification of sheetUnclassified.....

(As far as possible this sheet should contain only unclassified information. If it is necessary to enter classified information, the box concerned must be marked to indicate the classification eg (R) (C) or (S))

1. DRIC Reference (if known)	2. Originator's Reference Memorandum 3936	3. Agency Reference	4. Report Security U/C Classification	
5. Originator's Code (if known)	6. Originator (Corporate Author) Name and Location Royal Signals and Radar Establishment			
5a. Sponsoring Agency's Code (if known)	6a. Sponsoring Agency (Contract Authority) Name and Location			
7. Title A Pattern recognition approach to understanding the multi-level perceptron.				
7a. Title in Foreign Language (in the case of translations)				
7b. Presented at (for conference papers) Title, place and date of conference				
8. Author 1 Surname, initials Longstaff, I D	9(a) Author 2 Cross, J F	9(b) Authors 3,4...	10. Date	pp. ref.
11. Contract Number	12. Period	13. Project	14. Other Reference	
15. Distribution statement Unlimited				
Descriptors (or keywords)				
continue on separate piece of paper				
Abstract (Use Abstract)				

LMED
-8