Artificial Neural Networks
Lecture Notes

Stephen Lucci, PhD

Part 8

About this file:

- PThis is the printer-friendly version of the file "*lecture08.htm*". In case the page is not properly displayed, use IE 5 or higher.
- Since this is an offline page, each file "lecture*nn*.htm" (for instance "lecture08.htm") is accompanied with a "img*nn*" folder (for instance "img08") containing the images which make part of the notes. So, to see the images, Each html file must be kept in the same directory (folder) as its corresponding "img*nn*" folder.
- If you have trouble reading the contents of this file, or in case of transcription errors, email gi0062@bcmail.brooklyn.cuny.edu
- Acknowledgments:
Background image is from http://www.anatomy.usyd.edu.au/online/neuroanatomy/tutorial1/tutorial1.html (edited) at the University of Sydney Neuroanatomy web page. Mathematics symbols images are from metamath.org's GIF images for Math Symbols web page. Other image credits are given where noted, the remainder are native to this file.

Contents

Hidden Nodes As Feature Extractors

- What are the features in the following training set?

11 inputs

Table 6.1 Vectors for feature detection.

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	y
1	1	1	1	1	1	1	1	1	0	0	1
1	1	1	0	0	1	1	1	1	0	0	1
0	0	1	1	1	1	1	1	1	1	1	0
0	0	1	1	0	1	1	1	1	1	1	0

single output

?

How about X_3 and $X_6 \to X_9$

X_1 or X_{11} note correlation with outputs

- A feature: is a subset of the input space that helps us to discriminate the patterns by its examination without examining the entire pattern.

- Features contain the essential information content in the pattern set.

- If we trained a single semi-linear node using these vectors - we would expect:
    - A large positive weight to develop on input 1.
    - A negative weight on input 11.

- More formally - Use bipolar inputs and outputs
  $0 \longrightarrow -1$
  $1 \longrightarrow 1$
  We have $\bar{x}_i$, $\bar{y}$

- For each pattern and each component form $\bar{x}_i^p \, \bar{y}^p$ - measure of input/output correlation.

- Mean correlation C_i on the i^{th} component is

$$C_i = \sum_p \frac{1}{4} \bar{x}_i^p \, \bar{y}^p$$

- Plotting C_i vs. i

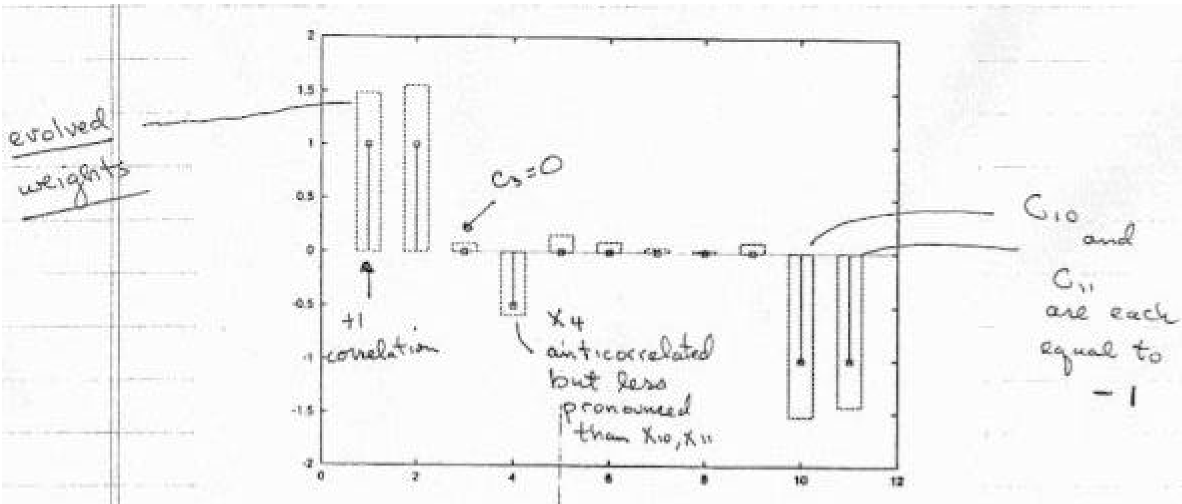


Figure 6.13 Weights and features.

Generalization and Overtraining

- In the left-hand graph (below), the training set has been "memorized" - poor generalization.
  The graph on the right has some errors but performs better on validation set.

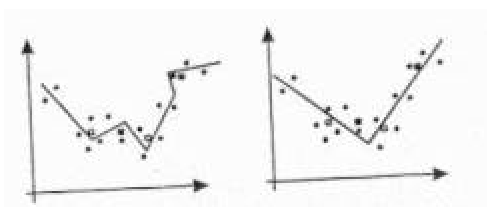


Figure 6.14 Generalization and overtraining in pattern space.

- Similar scenario:

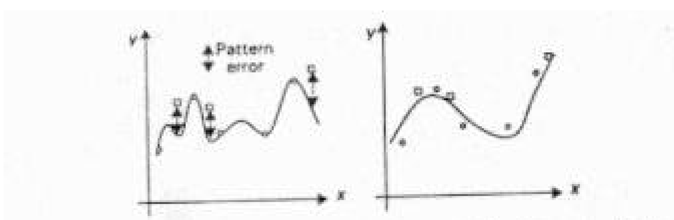


Figure 6.15 Generalization and overtraining in the context of function approximation.

- How can we avoid overtraining?

  Stop training when validation error is minimal.

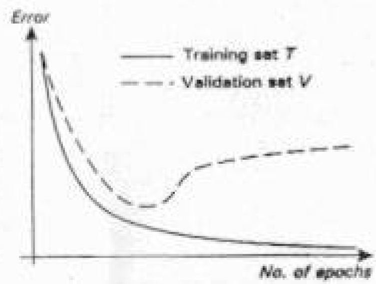


Figure 6.16 Cross-validation behaviour.

nonlinear regression. These similarities are explored further in the review article by Cheng & Titterington (1994).

On Data Representation

Internal Representation Issues

An example: *Credit Card Application*

Our database would consist of the these types of data:

- Numeric — income
expenses
number of current credit cards
- Category — Sex
Rent/Own
Checking Acct.
etc. ...
- Free-form Text — Occupation
Name
Address
etc. ...

- Name: Identifies the individual.
  Does not have a significant bearing on the credit-worthiness of the person.

- Address & Zipcode: Could be useful for integrating geographic block-code overlay data.
  N.B. It is illegal to use zipcode as an input by itself (redlining).

- Social Security Number: Could be useful as a key into other databases, but not directly useful here.

- Sex: Another field that cannot be used as network input!

- Marital Status: Can take on one of four values, typically encoded using a "one of N" coding - ie. one from among a set of N values.
  Example:
  M = Married; S = Single; W = Widowed; D = Divorced

For a Married subject we may have:

M	S	W	D
1	0	0	0

- Number of Children: May range from 0 to 9 (or more) - rescale into the range 0 to 1.

- Number of Children: Similarly as above, rescale the number into the range 0 to 1.

- Occupation: Some problems here ...
  More than 3,000 standard occupations and likely very few examples of each in our database.
  Therefore, group them into some larger set of categories: e.g.,
  Management
  Professional
  Skilled Labor
  Unskilled Labor
  etc. ...

- House Ownership: One of two categories - Rent or Own
  Therefore, "one of N" encoding, as we did for the Marital status entry.

- Monthly Income & Expenses: - Any suggestions ???

- Checking & Savings Accounts: Flags for Y = Yes, N = No.
  Thus,
  Y $\longrightarrow$ 1
  N $\longrightarrow$ 0
  or two inputs each ...(illegible.)

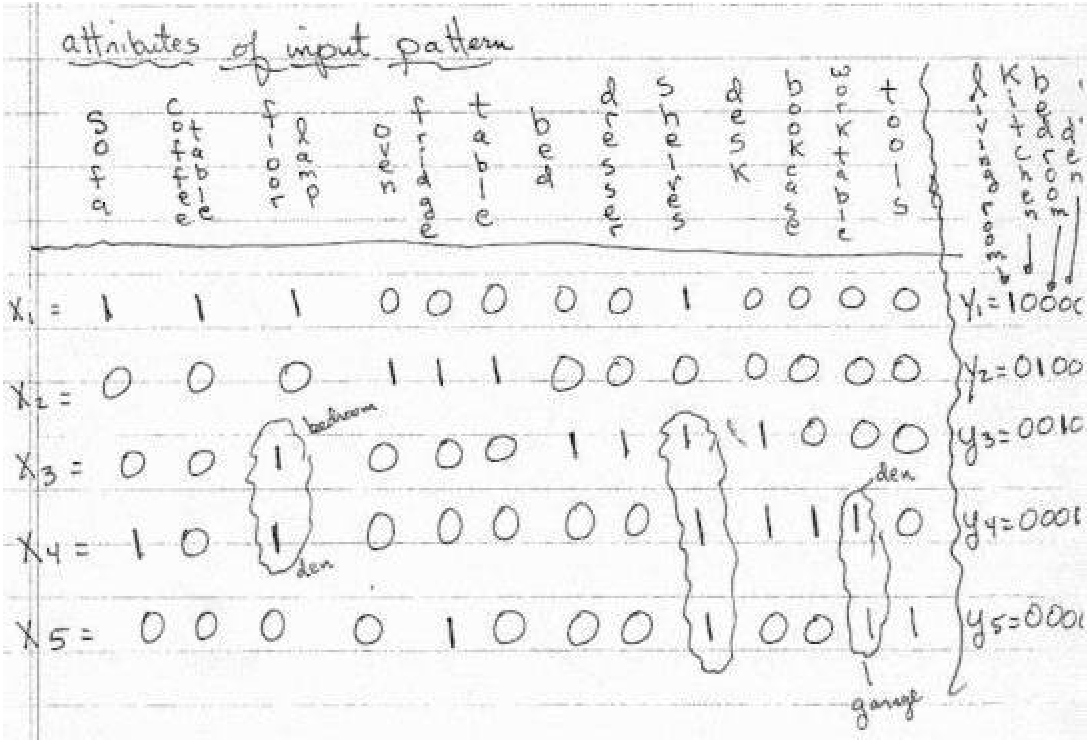
On Pattern Representation

- Neural network application to map a set of attributes to a room designation - room example revisited.

- Modify your previous data representation to include attributes that are common to more than one room.

- For example, a fireplace could be an attribute of a living room and a game room.

- Attributes of input pattern:



and what if the <u>attribute</u> is <u>unknown</u> for a certain room???

External Interpretation Issues
The processing elements within network have little or no understanding of the application.
Recall the L vs. $\mathscr{L}$ example.

Exemplar Analysis

- Accurate representation of a problem space.
- No inconsistencies.
- Problems can be corrected.

Ensuring Coverage
1:1 rule.
About half the inputs should be null patterns.

Exemplar Consistency Checking
Binary Search of exemplar set.

Resolving Inconsistencies

- <u>Eliminating Patterns</u> - if unlikely events.
- <u>Combining Patterns</u> - but be careful if outputs should be mutually exclusive.
- <u>Altering the Representation Scheme</u> - add a time or location stamp.

Training Guidelines

- Train BPN until global error falls below 0.2, note the epoch and save the state of the network for archival purposes.
- Continue training until error falls below 0.1
- If additional number of training epochs <= 30 % of original, then repeat process and try for 0.05.

After training - We test!

Partial-Set Training
Present the network with patterns not in the training set.

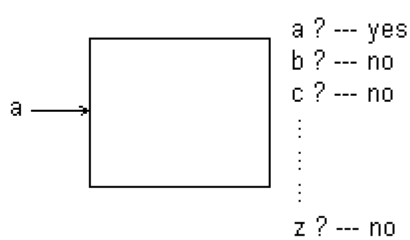
- Withold a number of patterns randomly
  However ... network may be sensitive to the order in which training patterns are presented.
- Hold-One-Out Training: Network must be trained n times. costly!

Pathology Analysis
Hinton diagrams allow developer to view activity of each unit in the network.

Applications in Pattern Classification

Pattern Classification
The final state of the network is interpreted in a manner that allows us to classify the input pattern as belonging to one of several categories.

Example:
Handwritten letter analysis



Handwritten letter analysis

Applications

- NETtalk - A "talking typewriter."
- Radar Signature Classifier - Aircraft identification

- Prostate Cancer Detection - Analysis of sonargrams.

NETtalk

Developed by Terry Sejnowski, Charles Rosenberg.

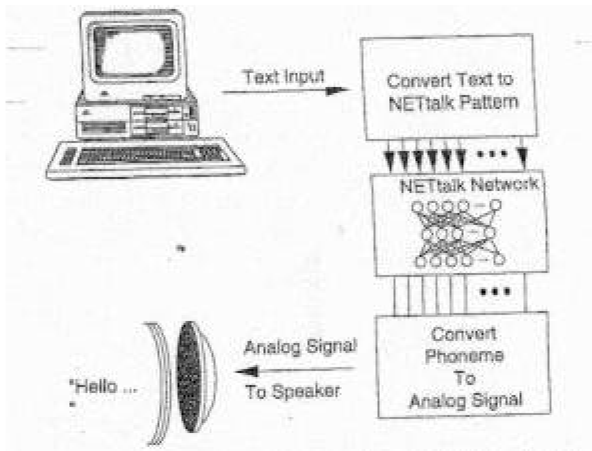


Figure 6.1 The system block diagram of the NETtalk application is shown. In this architecture, a front-end process converts the scanned text into the pattern representation for the BPN. The network, in turn, produces an output that is interpreted by a back-end process as the phoneme to be generated. The output of the system is therefore a synthetic pronunciation of written text.

A BPN is trained to classify a character sequence as one of 26 possible *phonemes* which are used to generate synthetic speech.

Pronounciation in English is problematic.
"Throw out the rules."
Consider

- *rough* vs. *through*
- *pizza* vs. *fizzy*
- *tortilla* vs. *villa*
- etc. ...

The pronounciation of the vowel(s) is dependent on a learned relationship between the vowel and its neighboring characters
(e.g., Note the *e* in *help* and *heave*.)

NETtalk captures the relationship between text and sounds by using a BPN to learn these relationships through experience.

>The textual representation of the word is converted into a pattern that the network can use.

Sejnowski and Rosenberg use a sliding-window for text.

The preceding three letters and the following three provide a <u>context</u> for a letter.

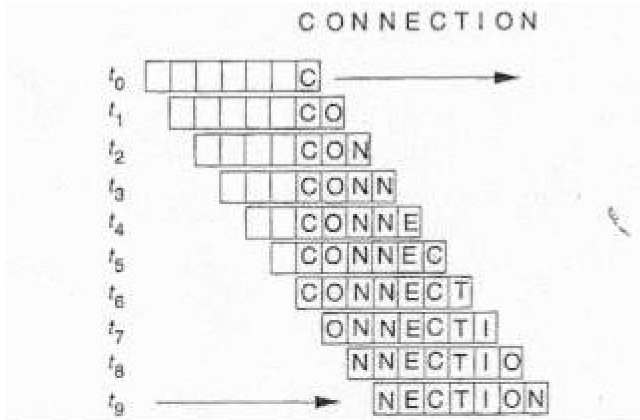


Figure 6.2 This diagram illustrates the sliding-window concept used to produce word patterns for the neural network. Initially, the window is padded with blanks to produce the silent phoneme. Characters from the word to be pronounced are then acquired by sliding the window over the word, one character at a time. This process continues until all of the characters in the word have passed by the focus position, with blank characters appended to the word to signal termination.

Letters are converted into pattern vectors consisting of 29 binary elements:

- 26 uppercase English alphabet characters.
- Three for punctuation characters that influence pronounciation
- "One of N" encoding (pattern characters are orthogonal.)

NETtalk Training

- <u>Training Data</u>:
  5000 common English words together with the corresponding phonetic sequence for each word.

- <u>Input Patterns</u>:
  203 binary inputs per pattern.
  (a 7-character window) * (29 elements/char).

- <u>Size of training set</u>:
  Average word length = 6 characters. Thus,
  (5000 words) * (6 char/word) = 30,000 exemplars.

NETtalk Architecture

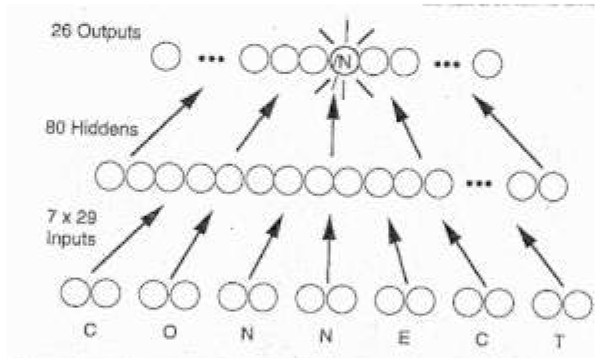


Figure 6.3 This diagram illustrates the architecture of the BPN used to perform the NETtalk application. The operation of the network is described in the text.

NETtalk Results
The classification produced by the network converted into the proper phoneme, and used to drive a speech synthesizer.

- Before Training
  The network produced random sounds, mixing consonant and vowel sounds.
- After 100 epochs
  The network began to separate words, recognizing the role of blank characters.
- After 500 epochs
  Clear distinction between vowel and consonant sounds
- After 1000 epochs
  Words distinguishible but not phonetically correct.
- After 1500 epochs
  The network captured phonetic rules - correct pronounciation, but mechanical sound.

• Training stopped and NETtalk was given 200 words to pronounce (obviously NOT from the training set

• NETtalk can read English text with an accuracy of "about 95%".
Hence, the network did not merely memorize a set of words and pronounciations. It learned the *general interrelations* between English text and sounds.

• NETtalk learning vs. a child learning to read.

Radar Signature Classifier

Pulse Doppler Radar - Uses

- Detecting airborne targets.
- Traffic law enforcement.

Determining *range* and *velocity* of target relative to the radar system.

Physical Principles and Problem Formulation

- Electromagnetic radiation (*EMR*) travels at a constant speed.
- EMR waves reflected from a moving body are frequency shifted in the direction of travel (The Doppler Effect, or "Red-Shift").

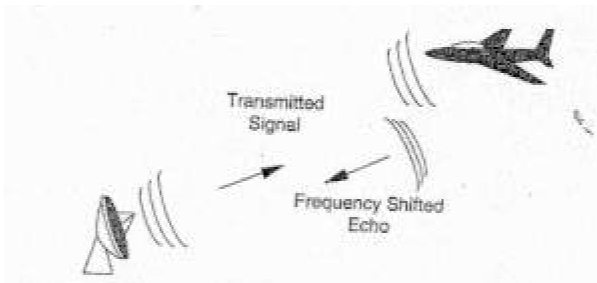


Figure 6.4 This diagram illustrates the basic principles that govern the operation of pulse Doppler radar. A radar pulse, or *chirp*, is transmitted from the radar station. The EMR burst travels outward from the transmitter at the speed of light. When an EMR-reflective target is hit by the outgoing signal, some of the energy is reflected back to the radar receiver, which then determines the range and velocity of the target based on concepts described in the text.

- Each radar target has its own unique way of reflecting radar emissions.

- Skillled radar operators are able to determine the type of target, even though the radar itself has no capability of making this determination.

  The straightforward approach:
  Use a fixed-frequency radar pulse.
  Determine *range* as a function of the delay between transmission and the reception of an *echo*.
  Determine *velocity* as a function of the *phase-shift* in the return frequency.

- Radio waves travel at a constant rates. The elapsed time between the transmitted and received signals provides the distance to the target. Therefore,

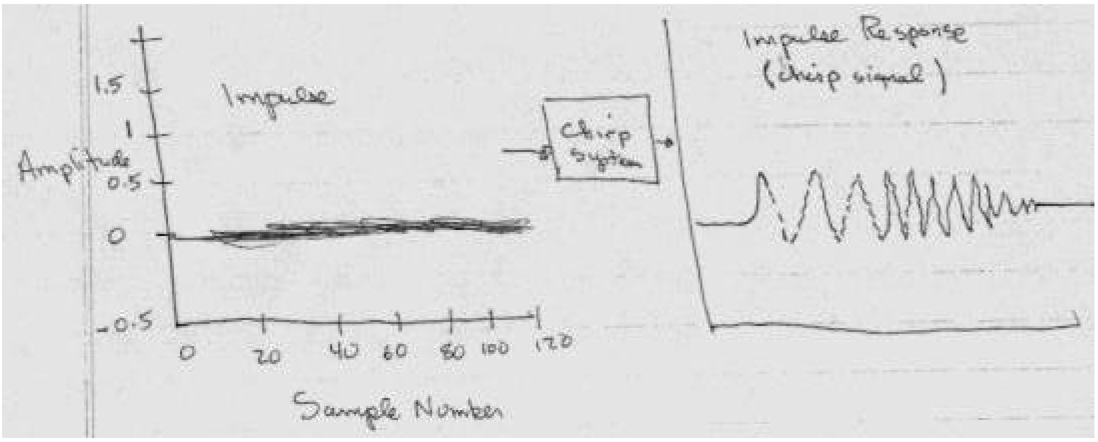
- *First Requirement for the pulse*:
  It needs to be as short as possible.

  Example:
  A 1 microsecond pulse provides a radio burst about 300 meters long.

- *Second Requirement for the pulse*:
  If we want to detect objects farther away, we need more energy in the pulse.
  Unfortunately, more energy and shorter pulse are conflicting requirements.

- ***Chirp signals***
  Chirp signals provide a way of breaking this limitation. Before the impulse reaches the final stage of the radio transmitter, it is passed through a *chirp system*.

- The reason it is called a chirp signal is because it sounds like the chirp of a bird, when played through a speaker.



- A *Fourier Transform* is used in processing.

- Inspection of the Doppler signature reveals information about the nature of the target.
  Example:
  If the airframe of the target has no moving blades (jet aircraft?) the Doppler signature is fairly compact.

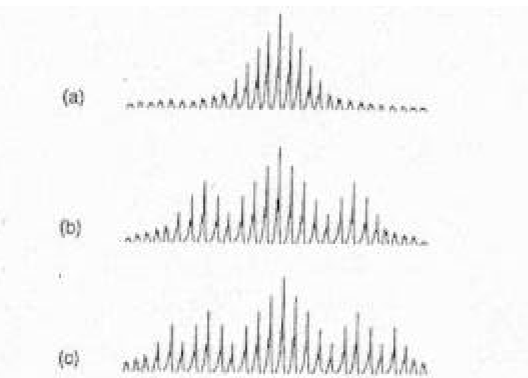


Figure 6.8 These diagrams illustrate Doppler signatures produced by different kinds of aircraft. (a) A jet (airframe only) produces a compact Doppler signature. (b) A two-bladed helicopter causes a spread in the signature, with side lobes forming on both sides of the peak, indicating rotation of the blades. (c) A four-bladed helicopter has a very wide signature, because blades rotate through the radar signal twice as often as with the two-bladed helicopter.

Learning
To Learn to classify these Doppler signatures, create a network that learns to produce the correct output for each signature.

<u>Problem</u>:
BPN is not a position-invariant pattern classifier.

Also, Doppler signatures shift as a function of target range:

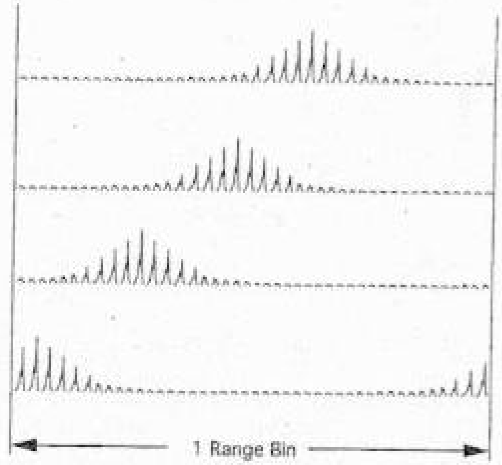


Figure 6.9 These graphs show how a Doppler signature shifts as a function of target range. Beginning at the top, the sequence shows a target approaching the radar station, with a corresponding left shift in the Doppler signature.

To compensate for the pattern shift, the network must be trained with exemplars that account for the pattern shift.

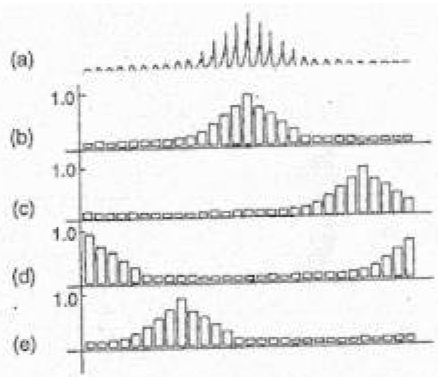


Figure 6.10 This diagram illustrates the exemplar encoding scheme used to capture the Doppler signature of various aircraft. (a) This diagram illustrates the general form of the radar return after extracting the signature. (b) The quantized form of the Doppler signature. (c) The quantized signature shifted right by ten range bins from the general form. (d) The quantized signature shifted by fifteen range bins. (e) The quantized signature shifted right by 26 range bins (or left by 6 range bins).

Architecture of the Network

- Three-layer BPN.
- 32 input units.
- 8 units on the hidden layer.
- 3 output units.

Training

- 96 Doppler signature aircraft-classification pairs.
- Learning rate n = 0.5
- Momentum α = 0.6
- 1700 epochs resulted in error of 0.01.

$\triangle w(n) = \alpha\delta^{j}(n)x(n) + \lambda w(n-1)$