# Neural Networks

Nethra Sambamoorthi, Ph.D

Jan 2003

CRMportals Inc.,
Nethra Sambamoorthi, Ph.D
Phone: 732-972-8969
Nethra@crmportals.com

# What? … Saying it Again in Different ways

- Artificial neural network (ANN) models are a collection of computational modeling methods for prediction of yes/no response (or multinomial classes) or prediction of continuous response (or multiple dependent responses) a la the decision making of collection of neurons in the brain
  - We will accept it as a way of modeling and learn from it to further the process of modeling activities



**CRMportals Inc.,**
**Nethra Sambamoorthi, Ph.D**
**Phone: 732-972-8969**
**Nethra@crmportals.com**

# What Is Covered?

- You will get an overview of methods.
  - Single layer ANN
  - Two layer ANN
  - Basics of training the network
  - Back-propagation
    - Single layer (actual calc)
    - Comments (other methods)
  - Statistical relationships and equivalent statistical jargons

The reference book name whose graphs have been used is missing as this is a old presentation made available through document retrieval research. When I locate the book it will be mentioned – apologies. A surprise git for any one who could identify the book name.

CRMportals Inc.,
Nethra Sambamoorthi, Ph.D
Phone: 732-972-8969
Nethra@crmportals.com

# A Great Reference Source

- http://zernike.uwinnipeg.ca/jargon - an article written by Warren Sarle of SAS Inc.,
- This will take you to a whole world of neural network related literature and its comparision to statisiticians jargon



**CRMportals Inc.,**
**Nethra Sambamoorthi, Ph.D**
**Phone: 732-972-8969**
**Nethra@crmportals.com**

# What? … Saying it Again in Different ways

- Artificial neural network (ANN) models are a collection of computational modeling methods for prediction of yes/no response (or multinomial classes) or prediction of continuous response (or multiple dependent responses) a la the decision making of collection of neurons in the brain
  - We will accept it as a way of modeling and learn from it to further the process of modeling activities



**CRMportals Inc.,**
**Nethra Sambamoorthi, Ph.D**
**Phone: 732-972-8969**
**Nethra@crmportals.com**

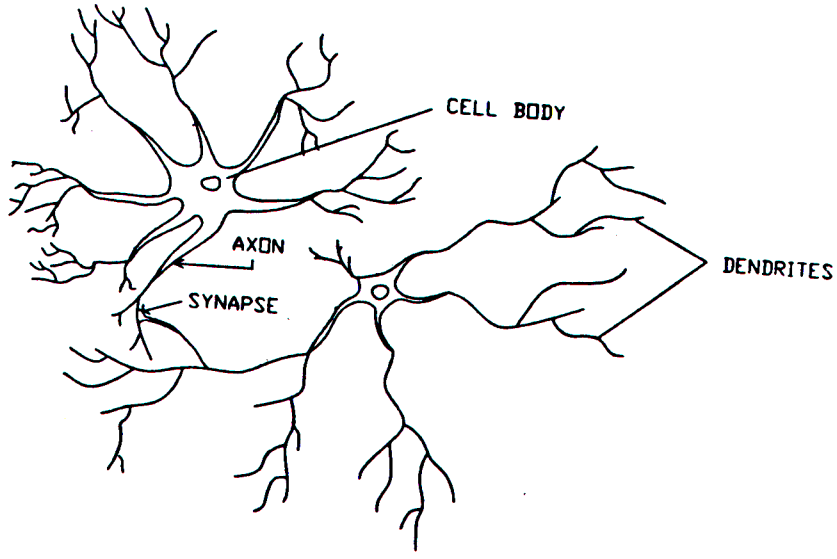
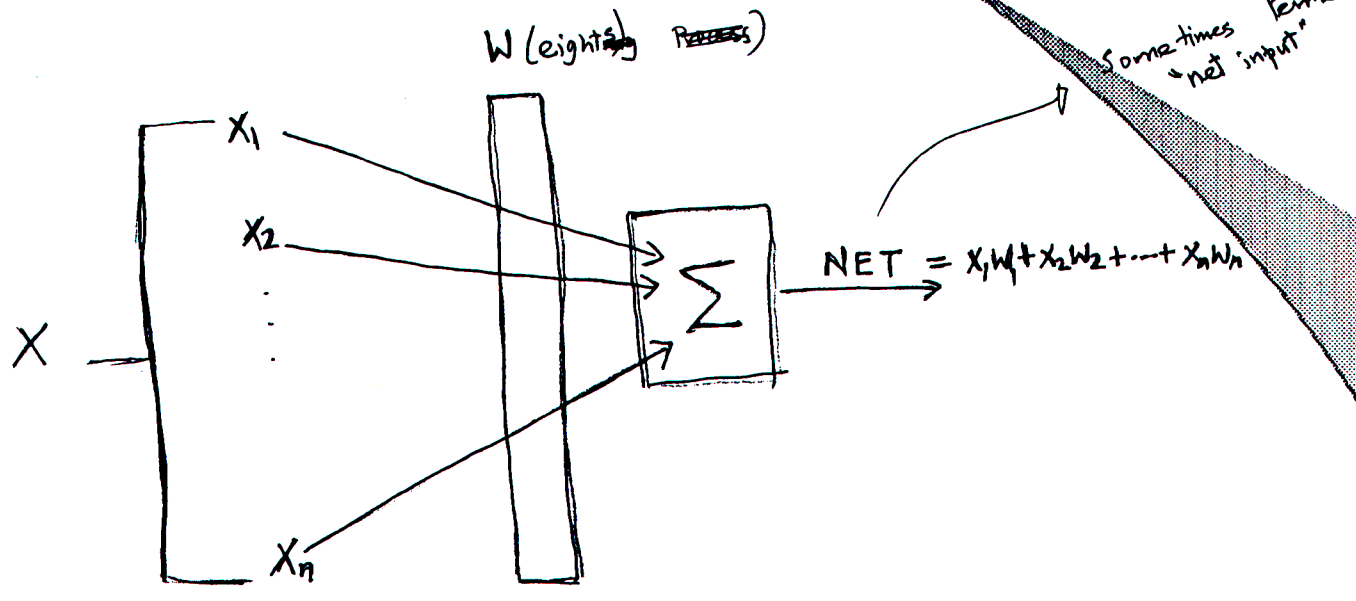



**Figure 1-1.** Biological Neuron

# Artificial Neuron

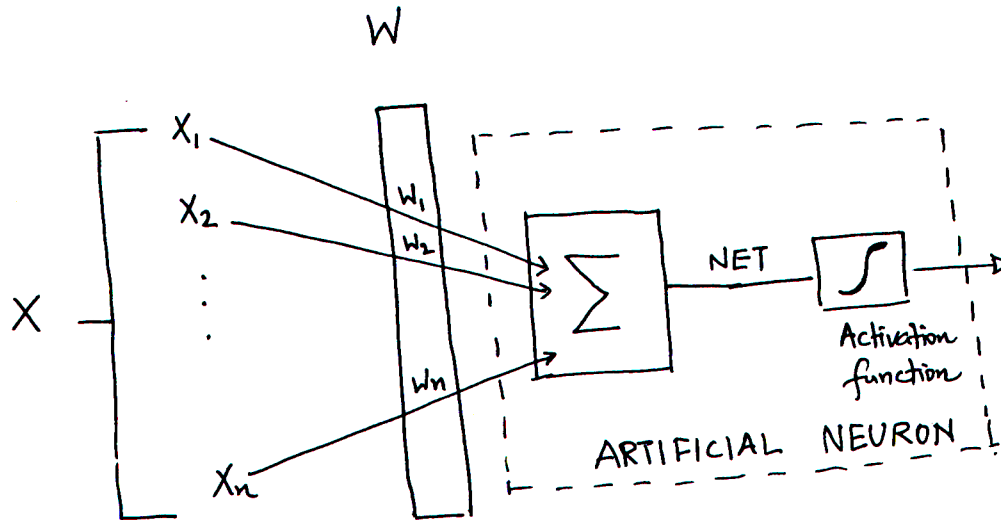



External inputs received by synapses

Strength of synaptic connection

Cell body (Neuron) (puts together the inputs) → the output of this 'putting together' is termed NET (Neuron)

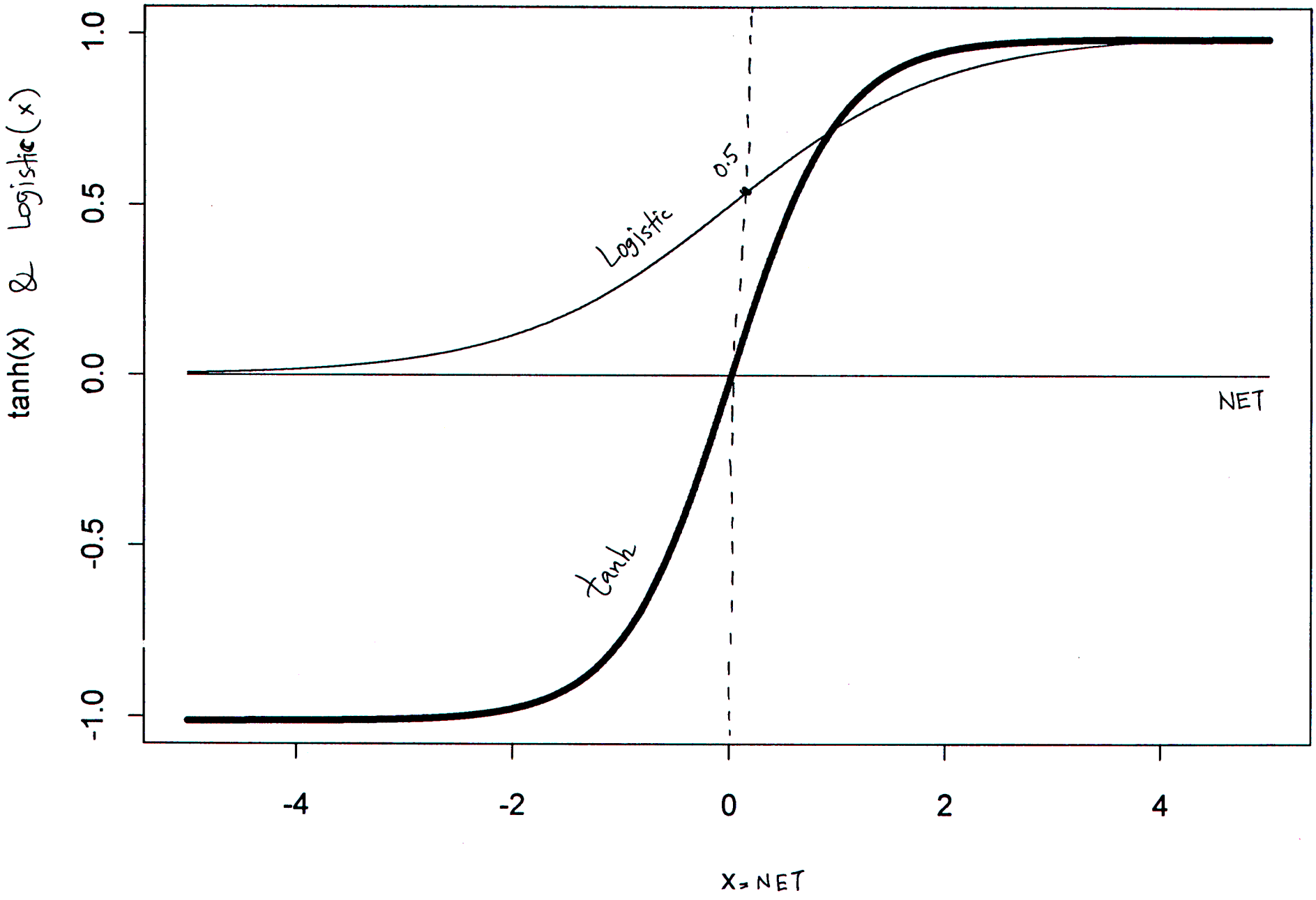
# Artificial Neuron with Activation Function




Usually logistic function if it is classification, tanh is another possibility

Sending / not sending a pulse (firing of a neuron) due to the significant / insignificant weight in the message inputs

# Comparison of Logistic and Tanh functions



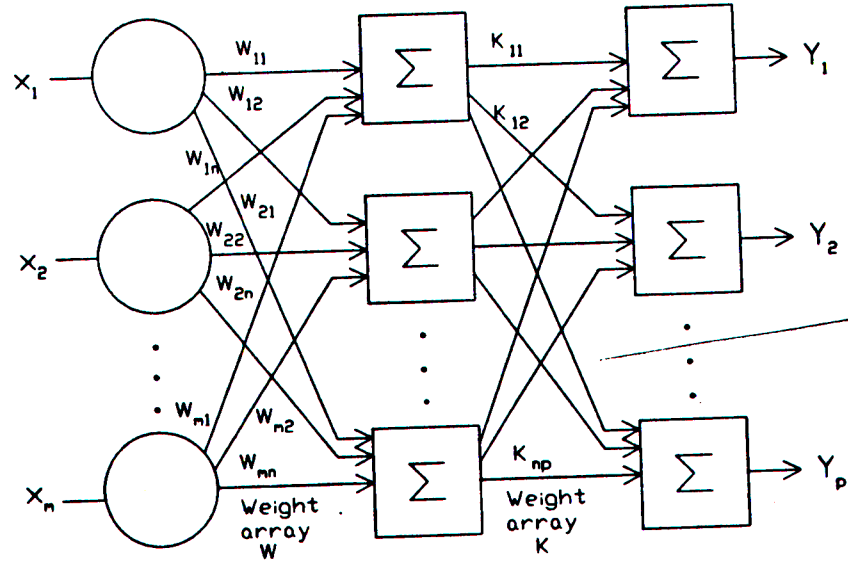



**Figure 1-6.** Two-Layer Neural Network

Example of a synaptic connection (algorithmic step connecting a cell to another cell)

Multiple cells = Multiple neurons

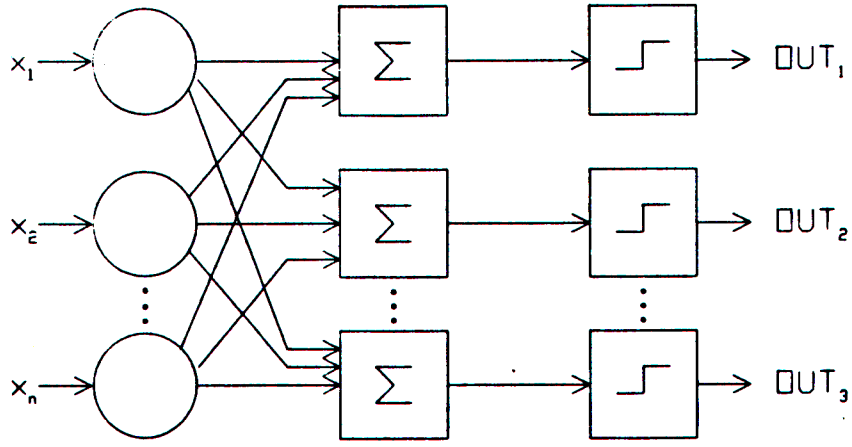



**Figure 2-2. Multioutput Perceptron**

# Back Propagation Steps

- Set the weights randomly before we begin the iteration process
  - Apply the first training sample point (calculation of weights is done using one sample point for one weighting coefficient at a time
  - Calculate the output of network

Forward pass

  - Calculate the error between network output and the target value for that sample
  - Adjust the weight of the vector to minimize the errors
  - Repeat the above steps for each input vector in the training set to adjust the weight coefficients until the error is acceptably low

reverse pass

CRMportals Inc.,
Nethra Sambamoorthi, Ph.D
Phone: 732-972-8969
Nethra@crmportals.com

# Training a network

(or)

How to find the 'optimal' weight vectors so that the network performs well with new data

- iterative procedure ; use one record for one co-efficient at a time

$$w_j^{(i+1)} = w_j^{(i)} + \underbrace{\Delta w_j^{(i+1)}}_{-\eta^{(i)} g_j^{(i)} + \alpha \Delta w_j^{(i)}}$$

(smoothing co-efficient) momentum term (0,1)

Step size (i)
- fixed in practice $\forall i$
- difficult to choose (0.1, 1)

partial derivative of error measure with respect to $w_j^{(i)}$

for logistic, $g_j^{(i)} = -F_j^{(i)}(1-F_j^{(i)})$ assuming RMS

- grdient descent method.

# A Great Reference Source

- http://zernike.uwinnipeg.ca/jargon - an article written by Warren Sarle of SAS Inc.,
- This will take you to a whole world of neural network related literature by statisiticians



**CRMportals Inc.,**
**Nethra Sambamoorthi, Ph.D**
**Phone: 732-972-8969**
**Nethra@crmportals.com**

# Neural Network and Statistical Jargon

Warren S. Sarle saswss@unx.sas.com Apr 29, 1996

URL: ftp://ftp.sas.com/pub/neural/jargon

The neural network (NN) and statistical literatures contain many of the same concepts but usually with different terminology. Sometimes the same term or acronym is used in both literatures but with different meanings. Only in very rare cases is the same term used with the same meaning, although some cross-fertilization is beginning to happen. Below is a list of such corresponding terms or definitions.

Particularly loose correspondences are marked by a ~ between the two columns. A < indicates that the term on the left is roughly a subset of the term on the right, and a > indicates the reverse. Terminology in both fields is often vague, so precise equivalences are not always possible. The list starts with some basic definitions.

There is disagreement in the NN literature on how to count layers. Some people count inputs as a layer and some don't. I specify the number of hidden layers instead. This is awkward but unambiguous.

| Definition | Statistical Jargon |
|---|---|
| generalizing from noisy data and assessment of the accuracy thereof | Statistical inference |
| the set of all cases one wants to be able to generalize to | Population |
| a function of the values in a population, such as the mean or a globally optimal synaptic weight | Parameter |
| a function of the values in a sample, such as the mean or a learned synaptic weight | Statistic |

| Neural Network Jargon | Definition |
|---|---|
| Neuron, neurode, unit, node, processing element | a simple linear or nonlinear computing element that accepts one or more inputs, computes a function thereof, and may direct the result to one or more other neurons |
| Neural networks | a class of flexible nonlinear regression and discriminant models, data reduction models, and nonlinear dynamical systems consisting of an often large number of neurons interconnected in often complex ways and often organized into layers |

| Neural Network Jargon | Statistical Jargon |
|---|---|
| Statistical methods | Linear regression and discriminant analysis, simulated annealing, random search |
| Architecture | Model |
| Training, Learning, Adaptation | Estimation, Model fitting, Optimization |
| Classification | Discriminant analysis |
| Mapping, Function approximation | Regression |
| Supervised learning | Regression, Discriminant analysis |
| Unsupervised learning, Self-organization | Principal components, Cluster analysis, Data reduction |
| Competitive learning | Cluster analysis |
| Hebbian learning, Cottrell/Munro/Zipser technique | Principal components |
| Training set | Sample, Construction sample |
| Test set, Validation set | Hold-out sample |
| Pattern, Vector, Example, Case | Observation, Case |
| Reflectance pattern | an observation normalized to sum to 1 |
| Binary(0/1), Bivalent or Bipolar(-1/1) | Binary, Dichotomous |
| Input | Independent variables, Predictors, Regressors, Explanatory variables, Carriers |
| Output | Predicted values |
| Forward propagation | Prediction |
| Training values Target values | Dependent variables, Responses, Observed values |
| Training pair | Observation containing both inputs and target values |
| Shift register, (Tapped) (time) delay (line), Input window | Lagged variable |
| Errors | Residuals |
| Noise | Error term |
| Generalization | Interpolation, Extrapolation, Prediction |

| | | |
|---|---|---|
| Error bars | | Confidence interval |
| Prediction | | Forecasting |
| Adaline (ADAptive LInear NEuron) | | Linear two-group discriminant analysis (not Fisher's but generic) |
| (No-hidden-layer) perceptron | ~ | Generalized linear model (GLIM) |
| Activation function, Signal function, Transfer function | > | Inverse link function in GLIM |
| Softmax | | Multiple logistic function |
| Squashing function | | bounded function with infinite domain |
| Semilinear function | | differentiable nondecreasing function |
| Phi-machine | | Linear model |
| Linear 1-hidden-layer perceptron | | Maximum redundancy analysis, Principal components of instrumental variables |
| 1-hidden-layer perceptron | ~ | Projection pursuit regression |
| Weights, Synaptic weights | < | (Regression) coefficients, Parameter estimates |
| Bias | ~ | Intercept |
| the difference between the expected value of a statistic and the corresponding true value (parameter) | | Bias |
| Shortcuts, Jumpers, Bypass connections, direct linear feedthrough (direct connections from input to output) | ~ | Main effects |
| Functional links | | Interaction terms or transformations |
| Second-order network | | Quadratic regression, Response-surface model |
| Higher-order network | | Polynomial regression, Linear model with interaction terms |
| Instar, Outstar | | iterative algorithms of doubtful convergence for approximating an arithmetic mean or centroid |
| Delta rule, adaline rule, Widrow-Hoff rule, LMS (Least Mean Squares) rule | | iterative algorithm of doubtful convergence for training a linear perceptron by least squares, similar to stochastic approximation |
| training by minimizing the median of the squared errors | | LMS (Least Median of Squares) |
| Generalized delta rule | | iterative algorithm of doubtful |

| | |
|---|---|
| | convergence for training a nonlinear perceptron by least squares, similar to stochastic approximation |
| Back propagation | Computation of derivatives for a multilayer perceptron and various algorithms (such as the generalized delta rule) based thereon |
| Weight decay, Regularization > | Shrinkage estimation, Ridge regression |
| Jitter | random noise added to the inputs to smooth the estimates |
| Growing, Pruning, Brain damage, Self-structuring, Ontogeny | Subset selection, Model selection, Pre-test estimation |
| Optimal brain surgeon | Wald test |
| LMS (Least mean squares) | OLS (Ordinary least squares) (see also "LMS rule" above) |
| Relative entropy, Cross entropy | Kullback-Leibler divergence |
| Evidence framework | Empirical Bayes estimation |
| OLS (Orthogonal least squares) | Forward stepwise regression |
| Probabilistic neural network | Kernel discriminant analysis |
| General regression neural network | Kernel regression |
| Topologically distributed encoding < | (Generalized) Additive model |
| Adaptive vector quantization | iterative algorithms of doubtful convergence for K-means cluster analysis |
| Adaptive Resonance Theory 2a ~ | Hartigan's leader algorithm |
| Learning vector quantization | a form of piecewise linear discriminant analysis using a preliminary cluster analysis |
| Counterpropagation | Regressogram based on k-means clusters |
| Encoding, Autoassociation | Dimensionality reduction (Independent and dependent variables are the same) |
| Heteroassociation | Regression, Discriminant analysis (Independent and dependent variables are different) |
| Epoch | Iteration |
| Continuous training, Incremental training, On-line training, Instantaneous training | Iteratively updating estimates one observation at a time via difference equations, as in stochastic approximation |

| | |
|---|---|
| Batch training, Off-line training | Iteratively updating estimates after each complete pass over the data as in most nonlinear regression algorithms |