Connectionist Models of Memory and Language

Edited by

Joseph P. Levy, Dimitrios Bairaktaris,
John A. Bullinaria, Paul Cairns

UCL
PRESS

CHAPTER 11

Neural networks: the new statistical models of mind

Nick Chater

Introduction

Neural network, connectionist or parallel distributed processing models of cognition have rapidly become dominant in many areas of cognitive science (e.g. McClelland & Rumelhart 1986, Rumelhart & McClelland 1986a, Gluck & Bower 1988, Seidenberg & McClelland 1989, Elman 1990, Hinton & Shallice 1991). Yet the scope and power of neural network models, and their relation to other approaches to modelling cognition, have been controversial (Fodor & Pylyshyn 1988, Pinker & Prince 1988, Fodor & McLaughlin 1990). At one extreme, there is a hope, frequently expressed by cognitive scientists informally but rarely put down in print, that neural network models will sweep away other approaches to modelling cognition, and in particular the symbolic models that have until recently dominated cognitive psychology and artificial intelligence. At the other extreme is the view that cognitive or psychological explanation is necessarily pitched at a symbolic level, and that neural networks are hence irrelevant to such explanation (Fodor & Pylyshyn 1988, but see Chater & Oaksford 1990). Advocates of this view argue that neural networks are simply a rediscovery of old-style statistical methods, with well known limitations, in reaction to which the symbolic model of mind (Fodor 1975, Newell & Simon 1976) was originally developed.

In this chapter, I review theoretical work which shows that there is a close relationship between various kinds of neural network and statistical models. This work has been developed within the technical literature on neural networks, but has not received wide attention within the cognitive modelling community. Within this literature, neural networks are viewed as statistical models, although they are models of a novel and powerful kind. The connection with the familiar territory of statistics helps to clarify the status and power of neural network models in cognitive science. It should not, I will argue, be taken to suggest that neural network models are simply reinventions of failed models of the past. Rather, I suggest they should be seen as a new development within a rich and varied history of statistical models of cognition. Furthermore, the connection with statistics helps clarify the relationship between neural network and symbolic models of cognition, and makes it clear that they have separate concerns, rather than standing in competition. I shall be concerned almost exclusively with *inferential* statistics as opposed to purely *descriptive* statistics (i.e. not statistics as mere collection of numbers, or as tools for conveniently displaying data).

The structure of this chapter is as follows. I begin by outlining the scope of statistics in very broad terms, stressing the generality of statistical methods. I then turn to the relationship between statistical methods and neural networks, concentrating on neural network learning methods, and dealing with supervised and unsupervised methods in turn. Finally, I draw conclusions for the place of neural network models in the history of psychology and their relationship with other modelling approaches, in particular the symbolic approach.

What is statistical inference?

The elements of probability theory and statistics (I shall sometimes use "statistics" to refer to both of these, but distinguish the two when context requires it) are familiar to researchers in cognitive psychology and the cognitive sciences generally. However, statistics are frequently encountered in their role as tools for data analysis, rather than in their broader context as method for inference. It is in this latter context that statistical methods can plausibly be viewed as models of cognition (and we shall consider some aspects of the psychological tradition of statistical modelling, in relation to neural network models below). Moreover, because of the dominance of a limited "data analysis" view of statistics in certain areas of the cognitive sciences, the claim that neural networks might be just statistical models is sometimes viewed with incredulity. Hence, we begin by sketching the broader view of statistics as very general mathematical methods for uncertain inference, within which statistical methods as used in data analysis in the cognitive sciences form only a small part.

Statistical inference is founded upon the mathematical theory of probability, and the distinct statistical traditions differ on how this theory is understood. The interpretation of probability theory has been controversial since its very beginnings. Nonetheless, the most usual early interpretation of probability theory was as a tool for formalizing rational thought concerning uncertain situations, such as gambling, insurance and the evaluation of court-room testimony (Gigerenzer et al. 1989). Indeed, the very choice of the word "probability", which referred to the degree to which a statement was supported by the evidence at hand, embodied this interpretation – that is, "probability" originally signified "rational degree of belief". Jakob Bernoulli explicitly endorsed this interpretation when he entitled his definitive book *Ars conjectandi,* or the *Art of conjecture* (Bernoulli 1713). This "subjectivist" conception ran through the eighteenth and into the nineteenth centuries (Daston 1988), frequently without clear distinctions being drawn between probability theory as a model of actual thought (or more usually, the thought of "rational", rather than common, people (Hacking 1990)) or as a set of normative canons prescribing how uncertain reasoning should occur. In a sense, then, early probability theory itself was viewed as a model of mind.

As the distinction between normative and descriptive models of thought became more firmly established, probability theory was primarily seen as having normative force, as characterizing rationality; whether or not people actually followed such normative dictates was seen as a secondary question. A wide variety of arguments that purport to show that individual degrees of beliefs should obey the laws of probability calculus have been developed, based on betting quotients and "Dutch book" arguments (Ramsey 1931, de Finetti 1937, Skyrms 1977), theories of preferences (Savage 1954), scoring rules (Lindley 1982) and derivation from minimal axioms (Good 1950, Cox 1961, Lucas 1970). Although each argument can be challenged individually, the fact that so many different lines of argument converge on the very same laws of probability has been taken as powerful evidence for the view that degrees of belief can be interpreted as probabilities (e.g. see Howson & Urbach (1989) and Earman (1992) for discussions). The suggestion that probability theory can be viewed as a normative theory of uncertain reasoning sets the bounds of probability theory much wider than the confines in which it is frequently encountered in introductory textbooks. According to this view, probability theory is not just concerned with reasoning about coins, dice and accident rates, but is a calculus for rational thought.

Many inferential problems concern the relationship between models or hypotheses, and observation or data. Some of these problems are concerned with inferring the probability of various kinds of observation, given that the structure of the underlying model is known. So, for example, the model might be a fair coin, and the question of interest might be the probability that 50 heads or more will be obtained in 200 throws. Statistical inference, by contrast, applies in the opposite direction, using observed data to infer the structure of the underlying model. For example, given the observation of 50 heads in 200 throws, assessing whether the coin is unbiased, what its likely bias might be, and with what confidence the bias can be estimated, all involve statistical inference, since observed data are used to infer aspects of the underlying model.

The problem of inductive or statistical inference is very general, and arises, in different guises, in a variety of domains. In epistemology and the philosophy of science, the problem is that of choosing the hypothesis or theory which is best supported by a given body of empirical observations: this is the problem of *induction*. A particular approach to statistics, the Bayesian approach, is by far the most well developed formal account of inductive reasoning (e.g. see Horwich 1982, Howson & Urbach 1989, Earman 1992). In the context of psychology, cognitive science and artificial intelligence, machine learning, pattern recognition and the study of neural networks, statistical inference corresponds to the problem of *learning* underlying structure from experience. It is with this broad sense of the scope of statistics in view that the claim that the mind is an intuitive statistician (Gigerenzer & Murray 1987), or that cognitive processes can be viewed as statistical processes, can be understood. The claim is not merely that the mind performs *t* tests or ANOVAs (although this has been proposed (Kelley 1967)). It is that the dictates of statistical theory concerning inductive inference are descriptive, not just prescriptive, regarding certain aspects of thought.

The project of characterizing statistics is complicated by the variety of different statistical schools, many of whose differences stem, as noted above, from different interpretations of the probability calculus. So far, we have considered the subjectivist interpretation, according to which probabilities are primarily interpreted as concerning rational updating of degrees of belief. This viewpoint sees no fundamental distinction between inference from beliefs about hypotheses to beliefs about data (the standard probabilistic case), and statistical inference in the reverse direction. Bayes (1764) showed that inference in the two directions can be related by a simple corollary of the axioms of probability:

$$P(H_j|D) = \frac{p(D|H_j)P(H_j)}{\sum_{i=1}^{n} P(D|H_i)P(H_i)} \qquad (11.1)$$

This result is the foundation of Bayesian statistics, which allows the probability of a model or hypothesis H_j given data D to be estimated, given the probability of the data given each possible model or hypothesis H_i, and the prior probability of each H_i. By the application of Bayes's theorem, the normal laws of probability can be used to infer how probable each of a range of hypotheses is, given a data set, simply by mechanical calculation. Notice that the denominator is the same whatever hypothesis is under consideration, and acts as a normalization factor which ensures that the probabilities $P(H_i|D)$ sum to 1. It is often treated as a constant, and Bayes's theorem is then expressed, as above, by stating that $P(H_i|D)$ is proportional to $P(H_i|D)P(H_i)$.

According to a subjectivist interpretation, the prior probability $P(H_j)$ can be interpreted as an initial degree of belief in the hypothesis H_j. But for alternative views of probability, such as the frequentist interpretation (according to which probabilities are the limits of relative frequencies of repeated events (e.g. Fisher 1922, von Mises 1939)) and objectivist interpretation (according to which probabilities are objective properties of the world (Mellor 1971)), it is difficult to see how any sense can be made of such probability statements. For this reason, among others, various alternatives to Bayesian statistics have since been derived. The principal alternative schools are those of Fisher (1956, 1970) and Neyman and Pearson (e.g. Neyman 1950), and most standard statistical tests within the behavioural sciences (e.g. the *t* test, the ANOVA, χ^2 test) were developed by these schools (though the standard discussion of such tests in introductory statistical textbooks frequently blends incompatible elements of these approaches together – see Gigerenzer et al. (1989)). We shall focus on Bayesian statistical methods henceforth, since it is these, and related methods, that most closely relate to neural network models. Furthermore, the subjectivist, Bayesian approach relates probability and statistics most directly to problems of belief updating, and hence has the most natural relation to cognitive processing.

At this level of generality, it should be clear that there is no limitation on the nature or complexity of the models (hypotheses, theories) that can be assessed using Bayesian statistics, aside from the fact that they must be well enough specified that the probability of each data outcome can be calculated given that the model holds. That is, hypotheses or theories must constitute probabilistic models. (In practice, of course, many hypotheses are not well enough specified for this to be possible, and additional assumptions must be made in order to fill out the hypothesis or theory into a full probabilistic model, but we shall not be concerned with this issue here.)

Probabilistic models include deterministic models, which specify their data with probability 1, and models which are defined in terms of symbolic structures (e.g. sets of grammar rules), and learning for such models can proceed according to standard Bayesian procedures. Bayesian methods can also be adapted to assess parametrized classes of model (e.g. straight lines versus quadratic models in curve fitting (e.g. Young 1977)).

While there is in principle no limitation on model complexity, performing the appropriate calculations may be extremely difficult, involving severe mathematical and computational problems. Hence, in practice, researchers have been forced to concentrate on relatively simple underlying models. For example, in the domain of language, statistical research has focused on very simple stochastic models such as hidden Markov models (e.g. Huang et al. (1990) – note that the parameters of hidden Markov models are generally trained by maximum likelihood estimation, a Fisherian, rather than a Bayesian, method; however, it may be viewed as a special case of the Bayesian approach in which priors are uniform), and this has been true even when considering areas of natural language where such models have been shown to be inadequate (Chomsky 1957). It is this practical limitation that has led to the claim that "probabilistic" or "statistical" models of language or some other aspect of cognition are not able to capture its true structure. Taken at face value, the claim makes no sense, since any adequately specified model can, in principle, be used in statistical inference, and hence there is really no such class as the class of statistical models. What is meant by the claim is, presumably, that the simple stochastic models considered in current statistical studies are not adequate.

This means that if neural networks turn out to be closely related to statistical methods this does not necessarily mean that they are inadequate to model particular cognitive phenomena – for they are a new kind of statistical model, and must be considered on their own terms.

Neural networks and statistics

Neal (1993: 475) succinctly sums up the connection between neural networks and probability theory and statistics viewing "neural networks as probabilistic models, and learning as statistical inference". Since many neural networks used in cognitive modelling are feedforward networks trained with back-propagation, I shall concentrate on this case, considering the two halves of the view in turn. I then briefly consider unsupervised learning networks.

Supervised learning

Neural network architectures as probabilistic models

First let us consider how a neural network can be viewed as a probabilistic model (I follow Neal's (1993) development). Consider a neural network (Fig. 11.1) which takes vectors of real valued numbers, $\boldsymbol{x}$, as input and produces real valued vectors, $\boldsymbol{y}$, as output. Assuming that the network is deterministic, i.e. that the same input always results in the same output, the network architecture defines a function f, where $\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{w})$, where $\boldsymbol{w} = (\mathrm{w}_1, \ldots, \mathrm{w}_m)$ denotes the vector of m weights in the network. Let us now suppose that the target output, z, is just $\boldsymbol{y}$ with the addition of Gaussian noise, of fixed standard deviation σ (other noise functions can, of course, be considered, but this is the simplest). Once the input is specified, the probability of the target outputs is specified by

$$P(z|\boldsymbol{x},\sigma) \propto \exp\left(-|z - f(\boldsymbol{x},\boldsymbol{w})|^2 / 2\sigma^2\right) \tag{11.2}$$

Thus, a neural network with a particular architecture and a given set of weights defines a probabilistic model: the output probabilities are fully specified given the input probabilities, in accordance with Equation 11.2. The neural network architecture (defined purely by the pattern of connections between nodes) thus defines a family of probabilistic models, parametrized by the weights $\boldsymbol{w}$ associated with the connections.

This formulation, while appropriate for modelling feedforward networks to be trained by back-propagation, is not, of course, a helpful way to analyze all supervised networks. For example, if the network has a stochastic dynamics, then the output of the network may itself be a probability distribution, rather than a particular deterministic state. For example, in the Boltzmann machine (Hinton & Sejnowski 1986) the goal is to produce an output probability distribution which models that observed during learning. In deterministic versions of stochastic dynamics, such as the deterministic Boltzmann machines (Peterson & Anderson 1987), real-valued output units are considered to denote the probabilities of each discrete binary output, rather than denoting a real-valued number.

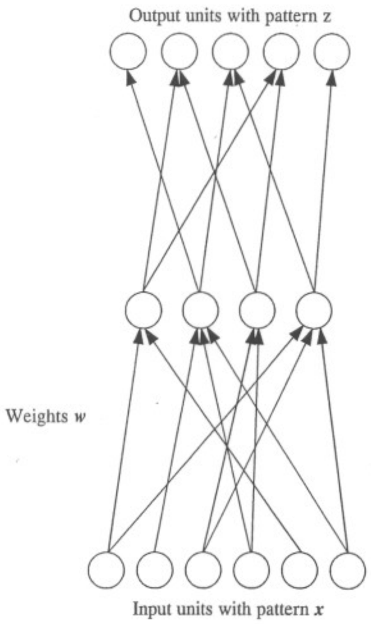

Figure 11.1 A supervised learning system. The inputs x_i are transformed into outputs y_i, which are compared against targets z_i. One natural goal of such reconstruction is to minimize $|z_i - y_i|^2$. In statistical terms, minimizing least squares can be viewed as assuming that there is Gaussian noise on the output. As noted in the text, there is sometimes an additional error term, which punishes networks with large weights $\boldsymbol{w}$.

Neural network learning as statistical inference

Before discussing the statistical interpretation, let us briefly summarize the back-propagation approach to training neural networks. Learning begins with a fixed network architecture and a specification of the target output, z_i, which is to be associated with each example input, $\boldsymbol{x}_i$. Back-propagation adjusts the weights in the light of these data. Typically, weights are adjusted so that some error function is minimized, the most common error function being the squared difference between the actual network output and the target output summed over units and patterns:

$$E(\boldsymbol{w}) = \sum_p |z_p - f(\boldsymbol{x},\boldsymbol{w})|^2 / 2\sigma^2 \tag{11.3}$$

In practice, it is sometimes found to be useful to allow weights to decay in proportion to their size, to discourage the network from developing extremely large weights, which sometimes lead to poor generalization. Hence the function to be minimized is modified to

$$E(\boldsymbol{w}) = \left(\sum_p |z_p - f(\boldsymbol{x},\boldsymbol{w})|^2 / 2\sigma^2\right) + \lambda|\boldsymbol{w}|^2 \tag{11.4}$$

where λ is a constant which sets the amount of "weight decay" used.

If the weights are adjusted in sufficiently small steps in the direction according to $-dE/dw_i$, the overall error E decreases. Eventually, the weights will reach a minimum, at which no local change decreases E. Unfortunately, there is no guarantee that this will be a global minimum, and so the network may not necessarily achieve the lowest possible E value. This problem of "local minima" is, however, a very general one, and typically applies in complex minimization problems which are solved iteratively (e.g. it arises in training hidden Markov models (Huang et al. 1990)).

The back-propagation learning algorithm is simply an efficient computational scheme for calculating the $-dE/dw_i$ values, by passing an "error" signal from the output units, where error is explicitly assigned, back through the rest of the network. It has the further advantage of being completely local – i.e. simple processes over the network units themselves serve to update the weights, and no external controller is required. From an abstract point of view, all that matters is that E is locally minimized somehow, and we shall not need to consider the details of back-propagation below.

A number of authors have shown how this learning algorithm can be viewed as statistical inference (e.g. Golden 1988, Buntine & Weigend 1991, Mackay 1992a, Neal 1992, Wolpert 1993). As noted above, we can consider the network and weight values to define a probabilistic model from which the data are considered to be generated, and aim to choose the weights which correspond to the most probable model, given the data $(x_1, z_1), \ldots, (x_n, z_n)$.

From Bayes's theorem it can be shown that

$$P\left(w|(x_1,z_1),\ldots,(x_n,z_n)\sigma\right) \propto P(w)P(z_1,\ldots,z_n|x_1,\ldots,x_n,\sigma,w) \qquad (11.5)$$

The probability of the data, given w and the assumption of Gaussian noise of standard deviation σ, is

$$P(z_1,\ldots,z_n|x_1,\ldots,x_n,\sigma,\omega) \propto \exp\left(-\sum_p |z_p - f(x,w)|_2 / 2\sigma^2\right) \qquad (11.6)$$

To calculate Equation 11.5 we must also specify some prior probability on w. (We assume here that the variance σ is known. Buntine & Weigend (1991) show that relatively minor modifications can deal with cases in which σ is unknown.) If we are interested in favouring small weights, then a natural prior is to assume that weight vectors are distributed in a Gaussian distribution, with standard deviation ω, around 0. That is,

$$P(w) \propto \exp\left(-|w|^2 / 2\omega^2\right) \qquad (11.7)$$

Substituting Equations 11.6 and 11.7 into Equation 11.5 gives

$$P\left(w|(x_1,z_1),\ldots,(x_n,z_n),\sigma\right) \propto \exp\left(-|w|^2/2\omega^2 - \sum_p |z_p - f(x,w)|^2/2\sigma^2\right) \qquad (11.8)$$

To maximize Equation 11.8 we minimize

$$E(w) = \sum_p |z_p - f(x,w)|^2 / 2\sigma^2 + |w|^2 / 2\omega^2 \qquad (11.9)$$

Thus we have a standard error function for back-propagation $E(w)$, as given in Equation 11.4, with $\lambda = 1/2\omega^2$. The parameter ω depends on the standard deviation of the Gaussian distribution of the priors. The smaller the standard deviation, the greater the bias towards networks with small weights.

Thus, we have a clear statistical interpretation of back-propagation learning. The strength of the weight decay term can now be understood in terms of how closely the prior distribution of weights is bunched around 0, i.e. it is determined by the value of ω. Interestingly, if the prior distribution is ignored, then the second term need not be considered, and we derive Equation 11.3. Thus, back-propagation without weight decay corresponds to computing the maximum likelihood weights, i.e. the weights according to which the data are most likely (Golden 1988).

A statistical interpretation of neural network performance is not just a mathematical curiosity. It makes sense of neural network learning, clarifies the assumptions underlying neural network performance, and provides insights into how neural network methods can be further developed. Thus, the use of least squares as a measure of error in statistical regression carries over as an appropriate measure of network error (the difference between the network's actual output and the specified output). In back-propagation networks, and variants, the weights are adjusted to perform gradient descent in this error. The statistical assumption underlying least squares is that the output value is subject to Gaussian noise; when this assumption is strongly violated, both statistical and neural network methods should ideally use an alternative error measure. For example, if the network outputs are known to be binary, an alternative measure, cross-entropy, is generally recommended as a more statistically appropriate error measure, and this has been widely used in neural network models. Here, then, statistics not only justifies standard methods, but suggests how they should be amended when necessary (see Hinton (1989) for discussion). Furthermore, a large range of new technical developments derive from the statistical interpretation (e.g. Mackay 1992a,b, Neal 1993, Wolpert 1993).

The statistical interpretation that we have considered amounts to viewing neural networks as a method for nonlinear regression, which is simply an extension of standard linear regression, which is a familiar data analysis tool in the behavioural sciences. Within conventional statistics, perhaps the most closely related approach to back-propagation is projection pursuit regression (Friedman & Stuetzle 1981), which has recently been related to neural network learning

(Intrator 1993). Standard linear regression aims to fit a straight line to a set of data points, so that least squares error is minimized, and this is justified as the maximum likelihood model, assuming Gaussian noise, just as in the network case described above. The analogue of weight decay in linear regression is systematically to favour lines with small regression components, and is known as ridge regression. The Bayesian analysis sketched above for neural networks with weight decay directly parallels a Bayesian rationale for ridge regression. Furthermore, linear regression is exactly modelled by a simplification of the standard back-propagation network – using no hidden units, and making the output units linear. Back-propagation affords a very considerable generalization over linear regression, since multilayered feedforward networks can learn to compute a very large class of nonlinear functions. Indeed, Hornik et al. (1989) have shown that any well behaved function can be approximated arbitrarily well by a neural network with sufficiently many hidden units.

Feedforward neural networks trained by back-propagation need not be viewed as a form of regression. With binary outputs, they can be viewed as classifiers, analogous to discriminant analysis. Indeed, with a single linear threshold unit (what Minsky & Papert (1969) termed a simple perceptron) they perform linear discriminant analysis between input points classified as 0 and input points classified as 1.

It is clear, then, that supervised networks, which are by far the most common network used in cognitive modelling, fit squarely in the tradition of conventional statistics, and are generalizations of familiar methods such as regression and discriminant analysis. We now turn to consider the statistical basis of unsupervised learning.

Unsupervised learning

Unsupervised learning methods involve finding structure in input data, with no specified "correct" output. The goal of the network is to extract interesting structure of some particular kind from the input. Unsupervised models have been much less used in modelling psychological data, although they have been viewed as a valuable source of hypotheses about aspects of human cognition (Kohonen 1984, Rumelhart & Zipser 1986, Ritter & Kohonen 1989, Finch & Chater 1992, Finch & Chater 1994; see also Ch. 12). An exception is Grossberg, who attempts to account for a large range of psychological data using rather elaborate unsupervised networks (e.g. Grossberg 1982). This work stands outside the mainstream of neural network research, and is beyond the scope of this chapter.

We shall briefly trace two connections between unsupervised learning and statistics. The first connection is simply that unsupervised learning methods frequently carry out identical or similar calculations to those of conventional statistical methods. For example, a one-layer feedforward network with lateral connections (Oja 1989) can learn to find principal components; competitive learning methods (such as that of Rumelhart & Zipser (1986)) can be viewed as computing slight variants of k-means cluster analysis (e.g. Krishnaiah & Kanal 1982).

The second connection has a deeper theoretical basis. Whereas supervised learning involves learning mapping between given input and target patterns, much unsupervised learning can be viewed as learning a mapping between input patterns and themselves – i.e. input and output are identical. In neural network terminology this is the "encoder" task. Since solving the encoder task is a special case of supervised learning, the statistical interpretation introduced above applies, and hence an appropriate error function is proportional to $|x_i - z_i|^2$, where z_i is the network output, and x_i is the pattern to be reconstructed.

The encoder task is trivial if a large enough network is available (in particular, when there are as many hidden units as input/output units) – the network can simply learn to perform the identity map. When the network is small, however, this is not possible, and to learn the task successfully the network must *compress* the input data into internal codes c_i, while losing as little information as possible. In order to compress data successfully, it is necessary to find structure within that data. To take a simple example, DeMers & Cottrell (1993) use standard back-propagation with a feedforward network to demonstrate that it is possible to compress input data which lie on a three-dimensional helix through a single hidden unit – thus, three-dimensional input data can be compressed onto a single dimension. In order to do this, the network must implicitly uncover the helical structure of the data, so that it can be represented by a single parameter. To take another example, Baldi & Hornik (1988) have shown that if a back-propagation network has just one hidden layer, then the units on that layer will extract the principal components of the input data (strictly, each of the n hidden units will find components which together span the subspace defined by the first n principal components, rather than finding exact principal components). It is because the goal of compression and reconstruction requires knowledge of the structure of the input that maximizing compression is an interesting goal of unsupervised learning. Indeed, there is also a direct theoretical connection between theoretical analysis of compression, in the minimum description length framework and Bayesian statistics (Rissanen 1983, Rissanen 1989), although I shall not consider this here.

In order to build a bridge between supervised and unsupervised learning, I have so far considered unsupervised learning methods which use standard feedforward networks trained by back-propagation. Of course, most unsupervised networks do not have this form; indeed, much interest in unsupervised learning concerns attempting to learn interesting structure without resorting to back-propagation and related methods. Nonetheless, the theoretical analysis sketched above can be used to derive many popular unsupervised learning algorithms.

Most unsupervised learning algorithms do not explicitly reconstruct the original input on a set of output units. Indeed, unsupervised networks generally consist only of input units, and what I shall call "feature" units which are

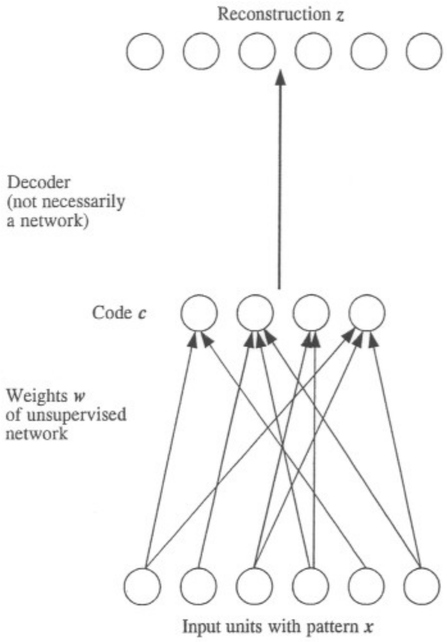
Reconstruction z
Decoder (not necessarily a network)
Code c
Weights w of unsupervised network
Input units with pattern x

Figure 11.2 An unsupervised learning system. The inputs x_i are transformed into code patterns c_i, which can be used to reconstruct the original input, giving z_i. One statistically natural goal of such reconstruction is to minimize $|z_i\text{-}x_i|^2$.

intended to display the structure implicit in the input. The above analysis can be applied by assuming a simple, fixed decoding mechanism, which maps feature unit patterns back onto the original input space (Fig. 11.2). Given this fixed decoding method, it is possible to calculate sum squared error as usual. Unsupervised algorithms can be viewed as adjusting their weights so that this implicit reconstruction can be as successful as possible - i.e. so that input data are compressed as well as possible.

For example, in competitive learning (Rumelhart & Zipser 1986) only a single feature unit is allowed to be active at any time, the unit whose weight vector is closest to the input pattern. The decoding mechanism for this network is simply to take the weight vector associated with the winning unit as indicating the input pattern. In order to minimize reconstruction error, it can be shown that the weights should move according to the standard competitive learning algorithm. Luttrell (1989, 1990, 1994) has extended this result to show that self-organizing maps similar (though not identical) to those of Kohonen (1984) can also be viewed as minimizing reconstruction error. Furthermore, networks which perform principal component analysis (Oja 1989) (without using full back-propagation) can also be understood in the same terms.

Discussion

I have outlined the close relationship between neural network models and statistics, and I now turn to considering the significance of this relationship for psychological theory. First, I shall attempt to put current neural network models in context in the history of psychology, arguing that they should be seen as descendants of previous statistical models of mind. Secondly, with the precursors of neural networks in mind, I shall draw out implications for the debate between neural network and symbolic approaches to cognition.

Relationship of neural networks to statistical models of mind

Neural networks are often portrayed as an entirely new and revolutionary approach to the mind (e.g. Clark 1989, Bechtel & Abrahamsen 1991); but by their critics they are frequently written off as associationism rediscovered (Fodor 1987, Fodor & Pylyshyn 1988). Neural networks do have close ties with a range of previous theories in psychology, including those based on associationist principles, although they are somewhat more complex in mathematical and computational terms. But they also have strong ties with a much broader tradition of modelling mental processes as involving statistical inference, and we shall briefly sketch some of these connections here (see Gigerenzer & Murray (1987) and Gigerenzer (1991) for further discussion of the tradition of statistical models of mind).

Perhaps the most well known statistical models have been outlined in the study of perception. The assumption that the mind makes psychophysical judgements and discriminations by using statistical techniques (based on Neyman–Pearson statistics) revolutionized psychophysics (Tanner & Swets 1954, Tanner 1965). The idea of the new "signal detection theory" was that the mind used statistical methods to take account of noise in perceptual stimuli. Earlier Brunswick (1943) had put forward a more general, but less mathematically sophisticated, doctrine of probabilistic functionalism, which held that mental statistical operations were necessary to integrate uncertain environmental cues. The methods of signal detection theory have since been applied to a broad range of cognitive processes, ranging from memory (Wickelgren & Norman 1966, Murdock 1982, Anderson & Milson 1989) to discriminating random from non-random patterns (Lopes 1981).

The study of similarity and categorization has also been influenced by statistical ideas. One statistically natural approach has been to model the environment as consisting of a number of distinct categories, which stochastically generate category examples. Given a particular category example, which must be classified, Bayes's theorem can be used to calculate the probability that it was generated by each of the possible categories. This approach to categorization gives rise to "likelihood" or "feature probability" models of categorization (Fried & Holyoak 1984, Anderson 1991). Thus, human categorization is viewed as involving the use of Bayesian statistics. Nosofsky (1990) has shown that the scope of this approach is actually rather wide, since it is mathematically extremely closely related to "exemplar" theories of categorization (Medin & Schaffer 1978, Nosofsky 1984, Estes 1986, Nosofsky 1986). Learning to categorize can itself be viewed as a (more difficult) problem of Bayesian inference, in which a particular number of types of generator category must be inferred. Recently, a psychological model of how this problem can be solved using Bayesian statistics has also been put forward (Anderson 1991). This kind of categorization model is formally closely related to mixture modelling approaches in neural networks (e.g. Jordan & Jacobs 1993).

Finally, statistical models have also been widely used in theorizing about human causal reasoning. For example, Kelley (1967) suggested that causal attribution was effected by conducting an intuitive ANOVA, and this approach has inspired a vast theoretical and experimental literature (e.g. see Cheng & Novick 1990).

The above discussion gives some idea of the breadth of the tradition of modelling mental processes in statistical terms, which stretches far beyond the confines of associationism. Hence, to suggest that neural networks lie within the tradition of statistical models of mind does not imply that they are simply a new form of associationism. Nonetheless, there are close connections between certain kinds of associative principle and particular neural network architectures. The best known relationship is, perhaps, that the Rescorla–Wagner law of classical conditioning is mathematically equivalent to the update rule for a single-layer neural network, one of the simplest neural network architectures (Gluck & Bower 1988). More sophisticated neural network-based models have also been used to attempt to provide new models of conditioning (Sutton & Barto 1981, Gluck et al. 1992).

Given the statistical interpretation of neural networks that we sketched above, neural network models can be viewed as lying firmly within this historical tradition of statistical models of cognition. But they do add something new. As I have argued, they add technical innovations so that the range of phenomena that can be modelled is much larger. Furthermore, neural networks are statistical algorithms implemented by a highly parallel processing architecture, which uses very simple processing units. Many, but by no means all, standard statistical methods can be efficiently implemented in this way; by implementing a statistical algorithm as a neural network we are automatically subject to an important constraint which appears to be a minimal condition for biological plausibility (Chater & Oaksford 1990).

Neural network models and symbolic theories of cognition

We are now in a position to reconsider the debate between symbolic and neural network approaches to cognition. As we noted above, advocates of neural networks sometimes argue that neural networks will entirely displace symbolic

models; and defenders of symbolic approaches to cognition have countered that neural networks are simply irrelevant to psychological explanation, which should be couched exclusively in symbolic terms (Fodor & Pylyshyn 1988).

According to the arguments presented here, this debate should really be cast in broader terms, as a debate between statistical and symbolic approaches to mind. But, once cast in these terms, the debate appears spurious, since the two approaches are concerned with orthogonal issues. The advocate of statistical methods pursues the possibility that aspects of cognition can be understood in terms of the apparatus of probability, statistics, information theory and decision theory. The advocate of symbolic methods pursues the claim that aspects of cognition involve the formal manipulation of structured symbolic representations (Fodor 1975, Newell & Simon 1976, Pylyshyn 1984). These are independent and entirely compatible claims about the nature of mind; they do not stand in competition. As we noted above, statistics tackles the problem of induction; but it does not place constraints on what is induced – it could be the grammar of a language, or an everyday or scientific hypothesis, all of which might be internally represented in symbolic form. If the debate between statistical and symbolic ideas seems ill conceived, the debate between neural networks (a special case of statistics) and symbolic ideas seems equally ill conceived.

I suspect that there has been a tendency to adopt a more radical view because of the difficulties which have been encountered in pursuing a symbolic approach to mind. Symbolic methods dominated the computational study of mind from the beginning of the cognitive revolution, with high expectations in certain quarters that the problems of human cognition might rapidly be unravelled. This optimism was based on the hope that early successes in formal domains, such as mathematical or logical reasoning, or game playing, should readily scale up to model common-sense thought. In practice, this symbolic program has run into serious obstacles, in capturing the densely interconnected and defeasible character of human knowledge and in devising mechanisms to reason with such knowledge (see Oaksford & Chater (1991, 1993) for extended treatments; see Dreyfus & Dreyfus (1986), Fodor (1983) and McDermott (1987) for related positions). The problems with modelling everyday thought have led to an increasing focus on apparently specialized cognitive processes, such as syntactic processing, early vision and motor control. But even here there have been considerable difficulties in attacking real-world problems – symbolic parsers cannot cope with natural text, and vision systems are very brittle when faced with real images.

From the point of view of psychology, this disarray does not offer an appealing menu of computational methods which can be recruited as the basis of potential cognitive models. It is therefore tempting to believe that neural networks offer a radical alternative paradigm, within which these difficulties either do not arise, or can readily be resolved. In fact, however, most neural network models are simply unable even to represent the problems that symbolic approaches were formulated to deal with, let alone solve them. There are, for example, no neural network models which parse real text, or analyze real visual scenes – even to begin to tackle such problems appears to presuppose the ability to represent complex structured information, for which symbolic representation is the only candidate. Instead of taking up the problems of the old symbolic paradigm and showing how they can be solved with neural networks, in practice, connectionist cognitive science has simply shifted focus onto different problems, which appear to be amenable to neural network analysis. Interest in neural network cognitive modelling has, for example, focused on highly specific domains such as reading (Seidenberg & McClelland 1989, Bullinaria 1993, Plaut & McClelland 1993), learning the past tense of verbs (Rumelhart & McClelland 1986b, Plunkett & Marchman 1991), finding structure in simple sequential material, and modelling aspects of speech perception and word recognition (e.g. McClelland & Elman 1986, Waibel et al. 1987, Abu-Bakar & Chater 1993, Cairns et al. 1994).

The rise of neural network models has not, in reality, been a revolution against old approaches, but simply a shift of emphasis away from one set of problems, which appear intractable, to another set of problems that can, perhaps, more readily be tackled. In particular, the problems that have been eschewed are just those in which structured representations are required. This picture is strengthened by the fact that those neural network models which have dealt with problems previously tackled by symbolic methods have done so not by overthrowing the symbolic approach, but by implementing symbolic structures and processes in terms of neural networks. So, for example, neural networks have been used to implement semantic networks (Hinton 1981, Shastri 1985, Smolensky 1987), production systems (Touretzky & Hinton 1985), schemata (Rumelhart et al. 1986) and specialist knowledge representation formalisms such as μ-klone (Derthick 1987).

I have argued for an ecumenical position: neural networks and statistics give rise to important tools for studying learning and uncertain reasoning; symbolic methods allow us to model the representation and processing of complex information. Both or neither of these approaches to the mind may ultimately prove fruitful; but there is no incompatibility between these approaches, and, for now, it seems appropriate to pursue both. From our current perspective, it is difficult to see how cognitive theory will be possible without making sense of both approaches, and showing how they can be integrated: the richness of symbolic representations and processes appears to be indispensable in processing language, in vision or in modelling everyday thought, and the statistical inductive methods which show how the information stored in such representations can be adjusted in the light of experience, appears to be equally indispensable. Both symbolic methods and neural networks are distressingly weak when viewed in the context of the extraordinary complexity of the real problems, in perception, language and common-sense thought, that people routinely and effortlessly solve. Cognitive science is, I would argue, currently better advised to develop and pursue both theoretical approaches, rather than to attempt to struggle along with either alone.

References

Abu-Bakar, M. & N. Chater 1993. Processing time-warped sequences using recurrent neural networks: Modelling rate-dependent factors in speech perception. *15th Annual Conference of the Cognitive Science Society, Proceedings*, 191–7. Hillsdale, New Jersey: Lawrence Erlbaum.

Anderson, J. R. 1991. The adaptive nature of human categorization. *Psychological Review* **98**, 409–29.

Anderson, J. R. & R. Milson 1989. Human memory: an adaptive perspective. *Psychological Review* **96**, 703–19.

Baldi, P. & K. Hornik 1988. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks* **2**, 53–8.

Bayes, T. 1764. An essay towards solving a problem in the doctrine of chances. *Royal Society of London, Philosophical Transactions* **53**, 370–418. Reprinted in *Biometrika* **45**, 296–315, 1958.

Bechtel, W. & A. Abrahamsen 1991. *Connectionism and the mind: an introduction to parallel distributed processing in networks*. Oxford: Oxford University Press.

Bernoulli, J. 1713. *Ars conjectandi*. Basel.

Brunswick, E. 1943. Organismic achievement and environmental probability. *Psychological Review* **50**, 255–72.

Bullinaria, J. A. 1993. Connectionist modelling of reading aloud. *2nd Workshop on the Cognitive Science of Natural Language Processing, Proceedings*. Dublin, 4–11.

Buntine, W. L. & A. S. Weigend 1991. Bayesian back-propagation. *Complex Systems* **5**, 603–43.

Cairns, P., R. Shillcock, N. Chater, J. Levy 1994. Lexical segmentation: the role of sequential statistics in supervised and un-supervised models. In *16th Annual Conference of the Cognitive Science Society*, A. Ram & K. Eiselt (eds), 136–41. Hillsdale, New Jersey: Lawrence Erlbaum.

Chater, N. & M. R. Oaksford 1990. Autonomy, implementation and cognitive architecture: a reply to Fodor and Pylyshyn. *Cognition* **34**, 93–107.

Cheng, P. W. & L. R. Novick 1990. A probabilistic contrast model of causal induction. *Journal of Personality and Social Psychology* **58**, 545–67.

Chomsky, N. 1957. *Syntactic structures*. The Hague: Mouton.

Clark, A. 1989. *Microcognition: philosophy, cognitive science and parallel distributed processing*. Cambridge, Mass.: Bradford Books/MIT Press.

Cox, R. T. 1961. *The algebra of probable inference*. Baltimore: The Johns Hopkins University Press.

Daston, L. 1988. *Classical probability in the enlightenment*. Princeton, New Jersey: Princeton University Press.

de Finetti, B. 1937. Foresight: its logical laws, its subjective sources. Translated in H. E. Kyburg & H. E. Smokler 1964 (eds), *Studies in subjective probability*. Chichester: John Wiley.

DeMers, D. & G. Cottrell 1993. Non-linear dimensionality reduction. See Hanson et al. (1993), 3–10.

Derthick, M. 1987. *A connectionist architecture for representing and reasoning about structured knowledge*. Department of Computer Science, Carnegie-Mellon University, Technical Report CMU-BOLTZ-29.

Dreyfus, H. L. & S. E. Dreyfus 1986. *Mind over machine: the power of human intuition and expertise in the era of the computer*. New York: The Free Press.

Earman, J. 1992. *Bayes or bust? A critical examination of Bayesian confirmation theory*. Cambridge, Mass.: Bradford Books/MIT Press.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* **14,** 179–211.

Estes, W. K. 1986. Array models for category learning. *Cognitive Psychology* **18**, 500–549.

Finch, S. & Chater, N. 1992. Learning syntactic categories using a neural network. *14th Annual Conference of the Cognitive Science Society, Proceedings*, 820–25. Hillsdale, New Jersey: Lawrence Erlbaum.

Finch, S. & Chater, N. 1994. Learning syntactic categories: a statistical approach. In *Neurodynamics and psychology*, G. D. A. Brown & M. Oaksford (eds), 294–321. London: Academic Press.

Fisher, R. A. 1922. On the mathematical foundations of theoretical statistics. *Royal Society of London, Philosophical Transactions A* **222**, 309–68.

Fisher, R. A. 1956. *Statistical methods and statistical inference*. Edinburgh: Oliver and Boyd.

Fisher, R. A. 1970. *Statistical methods for research workers*, 14th edn. Edinburgh: Oliver and Boyd.

Fodor, J. A. 1975. *The language of thought*. New York: Thomas Crowell.

Fodor, J. A. 1983. *The modularity of mind*. Cambridge, Mass: MIT Press.

Fodor, J. A. 1987. *Psychosemantics*. Cambridge, Mass.: Bradford Books/MIT Press.

Fodor, J. A. & B. P. McLaughlin 1990. Connectionism and the problem of systematicity: why Smolensky's solution doesn't work. *Cognition* **35**, 183–204.

Fodor, J. A. & Z. W. Pylyshyn 1988. Connectionism and cognitive architecture: a critical analysis. *Cognition* **28**, 3–71.

Fried, L. S. & K. J. Holyoak 1984. Induction of category distributions: a framework for classification learning. *Journal of Experimental Psychology: Learning, Memory and Cognition* **10**, 234–57.

Friedman, J. H. & W. Stuetzle 1981. Projection pursuit regression. *Journal of the American Statistical Association* **76**, 817–23.

Gigerenzer, G. 1991. From tools to theories: a heuristic of discovery in cognitive psychology. *Psychological Review* **98**, 254–67.

Gigerenzer, G. & D. J. Murray 1987. *Cognition as intuitive statistics*. Hillsdale, New Jersey: Lawrence Erlbaum.

Gigerenzer, G., Z. Swijtink, T. Porter, L. Daston, J. Beatty, L. Krüger 1989. *The empire of chance: how probability changed science and everyday life*. Cambridge: Cambridge University Press.

Gluck, M. A. & G. H. Bower 1988. From conditioning to category learning: an adaptive network model. *Journal of Experimental Psychology: General* **117**, 227–47.

Gluck, M. A., P. T. Glauthier, R. S. Sutton 1992. Adaptation of cue-specific learning rates in network models of human category learning. In *14th Annual Conference of the Cognitive Science Society, Proceedings*, 540–45. Hillsdale, New Jersey: Lawrence Erlbaum.

Golden, R. M. 1988. A unified framework for connectionist system. *Biological Cybernetics* **59**, 109–20.

Good, I. J. 1950. *Probability and the weighting of evidence*. London: Griffin.

Grossberg, S. 1982. *Studies of mind and brain: neural principles of learning, perception, development, cognition and motor control*. Boston: Reidell Press.

Hacking, I. 1990. *The taming of chance*. Cambridge: Cambridge University Press.

Hanson, S. J., J. D. Cowan, C. Lee Giles (eds) 1993. *Advances in neural information processing systems 5*. San Mateo, Calif.: Morgan Kaufman.

Hinton, G. E. 1981. Implementing semantic networks in parallel hardware. In *Parallel models of associative memory*, G. E. Hinton & J. A. Anderson (eds), 161–87. Hillsdale, New Jersey: Lawrence Erlbaum.

Hinton, G. E. 1989. Connectionist learning procedures. *Artificial Intelligence* **40**, 185–234.

Hinton, G. E. & T. J. Sejnowski 1986. Learning and relearning in Boltzmann machines. In *Parallel distributed processing: explorations in the microstructure of cognition*, vol. 1. *Foundations*, D. Rumelhart & J. L. McClelland (eds), 282–317. Cambridge, Mass.: MIT Press.

Hinton, G. E. & T. Shallice 1991. Lesioning an attractor network: investigations of acquired dyslexia. *Psychological Review* **98**, 74–95.

Hornik, K., M. Stinchcombe, H. White 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* **2,** 359–66.

Horwich, P. 1982. *Probability and evidence*. Cambridge: Cambridge University Press.

Howson, C. & P. Urbach 1989. *Scientific reasoning: the Bayesian approach*. La Salle: Open Court.

Huang, X. D., Y. Ariki, M. A. Jack 1990. *Hidden Markov models for speech recognition*. Edinburgh: Edinburgh University Press.

Intrator, N. 1993. On the use of projection pursuit constraints for training neural networks. See Hanson et al. (1993), 15–20.

Jordan, M. I. & R. A. Jacobs 1993. Hierarchies of adaptive experts. See Hanson et al. (1993), 985–92.

Kelley, H. H. 1967. Attribution theory in social psychology. In *Nebraska symposium on motivation*, vol. 1, D. Levine (ed.), 192–238. Lincoln: University of Nebraska Press.

Kohonen, T. 1984. *Self-organization and associative memory*. Berlin: Springer-Verlag.

Krishnaiah, P. R. & L. N. Kanal (eds) 1982. Classification, pattern recognition and reduction of dimensionality. *Handbook of statistics*, vol. 2. Amsterdam: North-Holland.

Lindley, D. V. 1982. Scoring rules and the inevitability of probability. *International Statistical Review* **50**, 1–26.

Lopes, L. L. 1981. Decision making in the short run. *Journal of Experimental Psychology: Human Learning and Memory* **8**, 626–36.

Lucas, J. R. 1970. *The concept of probability*. Oxford: Oxford University Press.

Luttrell, S. P. 1989. Self-organisation: a derivation from first principles of a class of learning algorithms. *3rd IEEE International Joint Conference on Neural Networks, Proceedings*. Washington, DC, vol. 2, 495–8.

Luttrell, S. P. 1990. Derivation of a class of training algorithms. *IEEE Transactions on Neural Networks* **1**, 229–32.

Luttrell, S. P. 1994. A Bayesian analysis of self-organising maps. *Neural Computation* **6**, 767–94.

Mackay, D. J. C. 1992a. A practical Bayesian framework for backpropagation networks. *Neural Computation* **4**, 448–72.

Mackay, D. J. C. 1992b. The evidence framework applied to classification networks. *Neural Computation* **4**, 698–714.

McClelland, J. L. & J. L. Elman 1986. Interactive processes in speech perception: the TRACE model. See McClelland & Rumelhart(1986), 58–121.

McClelland, J. L. & D. E. Rumelhart (eds) 1986. *Parallel distributed processing: explorations in the microstructures of cognition*, vol. 2. *Psychological and biological models*. Cambridge, Mass.: MIT Press.

McDermott, D. 1987. A critique of pure reason. *Computational Intelligence* **3**, 151–60.

Medin, D. L. & M. M. Schaffer 1978. Context theory of classification learning. *Psychological Review* **85**, 207–38.

Mellor, D. H. 1971. *The matter of chance*. Cambridge: Cambridge University Press.

Minsky, M. & S. Papert 1969. *Perceptrons: an introduction to computational geometry*. Cambridge, Mass.: MIT Press.

Murdock Jr, B. B. 1982. A theory for the storage and retrieval of item and associative information. *Psychological Review* **89**, 609–26.

Neal, R. M. 1992. *Bayesian training of backpropagation networks by the hybrid Monte Carlo method*. Department of Computer Science, University of Toronto, Technical Report CRG-TR-92-1.

Neal, R. M. 1993. Bayesian learning via stochastic dynamics. See Hanson et al. (1993), 475–82.

Newell, A. & H. A. Simon 1976. Computer science as empirical enquiry. *Communications of the ACM* **19**, 113–26. Reprinted in M. Boden (ed.), *The philosophy of artificial intelligence*. Oxford: Oxford University Press, 1990.

Neyman, J. 1950. *Probability and statistics*. New York: Holt.

Nosofsky, R. M. 1984. Choice, similarity and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory and Cognition* **10**, 104–14.

Nosofsky, R. M. 1986. Attention, similarity and the identification–categorization relationship. *Journal of Experimental Psychology: General* **115**, 39–57.

Nosofsky, R. M. 1990. Relations between exemplar-similarity and likelihood models of classification. *Journal of Mathematical Psychology* **34**, 393–418.

Oaksford, M. R. & N. Chater 1991. Against logicist cognitive science. *Mind and Language* **6**, 1–38.

Oaksford, M. R. & N. Chater 1993. Reasoning theories and bounded rationality. In *Rationality*, K. Manktelow & D. Over (eds), 31–60. London: Routledge.

Oja, E. 1989. Neural networks, principal components and subspaces. *International Journal of Neural Systems* **1**, 61–8.

Peterson, C. & J. R. Anderson 1987. A mean field learning algorithm for neural networks. *Complex Systems* **1**, 995–1019.

Pinker, S. & A. Prince 1988. On language and connectionism: analysis of a parallel distributed model of language acquisition. *Cognition* **28**, 73–193.

Plaut, D. C. & J. L. McClelland 1993. Generalization with componential attractors: Word and non-word reading in an attractor network. In *15th Annual Conference of the Cognitive Science Society, Proceedings*, 824–29. Hillsdale, New Jersey: Lawrence Erlbaum.

Plunkett, K. & V. Marchman 1991. U-shaped learning and frequency effects in a multilayered perceptron: implications for child language acquisition. *Cognition* **38**, 43–102.

Pylyshyn, Z. W. 1984. *Computation and cognition: toward a foundation for cognitive science*. Cambridge, Mass: Bradford Books/MIT Press.

Ramsey, F. P. 1931. *The foundations of mathematics and other logical essays*. London: Routledge and Kegan Paul.

Rissanen, J. 1983. A universal prior for integers and estimation by minimal description length. *Annals of Statistics* **11**, 416–31.

Rissanen, J. 1989. *Stochastic complexity in statistics inquiry*. Singapore: World Scientific.

Ritter, H. & T. Kohonen 1989. Self-organizing semantical maps. *Biological Cybernetics* **61**, 241–54.

Rumelhart, D. E. & J. L. McClelland (eds) 1986a. *Parallel distributed processing: explorations in the microstructures of cognition*, vol. 1. *Foundations*. Cambridge, Mass.: MIT Press.

Rumelhart, D. E. & J. L. McClelland 1986b. On learning the past tenses of English verbs. See McClelland & Rumelhart (1986), 216–71.

Rumelhart, D. E., P. Smolensky, J. L. McClelland, G. E. Hinton 1986. Schemata and sequential thought processes in PDP models. See McClelland & Rumelhart (1986), 7–57.

Rumelhart, D. E. & D. Zipser 1986. Feature discovery by competitive learning. In *Parallel distributed processing*, vol. 1. *Foundations*. D. E. Rumelhart & J. L. McClelland (eds), 151–93. Cambridge, Mass: MIT Press.

Savage, L. J. 1954. *The foundations of statistics*. New York: John Wiley.

Seidenberg, M. S. & McClelland, J. L. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review* **96**, 523–68.

Shastri, L. 1985. Evidential reasoning in semantic networks: a formal theory and its parallel implementation. Department of Computer Science, University of Rochester, Report TR166.

Skyrms, B. 1977. *Choice and chance*. Belmont: Wadsworth.

Smolensky, P. 1987. *On variable binding and the representation of symbolic structures in connectionist systems*. Department of Computer Science, University of Colorado at Boulder, Technical Report CU-CS-355-87.

Sutton, R. S. & A. G. Barto 1981. Towards a modern theory of adaptive networks: expectation and prediction. *Psychological Review* **88**, 135–70.

Tanner Jr, W. P. 1965. *Statistical decision processes in detection and recognition*. Sensory Intelligence Laboratory, Department of Psychology, University of Michigan, Technical Report.

Tanner Jr, W. P. & J. A. Swets 1954. A decision-making theory of visual detection. *Psychological Review* **61**, 401–9.

Touretzky, D. S. & G. E. Hinton 1985. Symbols among the neurons: details of a connectionist inference architecture. *9th International Joint Conference on Artificial Intelligence, Proceedings*, 238–43.

von Mises, R. 1939. *Probability, statistics and truth*. London: Allen Unwin.

Waibel, A., T. Hanazawa, G. E. Hinton, K. Shikano, K. Lang 1987. *Phoneme recognition using time-delay neural networks*. ATR Interpreting Telephony Research Laboratories, Japan, Technical Report TR-1-0006.

Wickelgren, W. A. & D. A. Norman 1966. Strength models and serial position in short-term recognition memory. *Journal of Mathematical Psychology* **3**, 316–47.

Wolpert, D. H. 1993. On the use of evidence in neural networks. See Hanson et al. (1993), 539–46.

Young, A. S. 1977. A Bayesian approach to prediction using polynomials. *Biometrika* **64**, 309–17.