Neural Computing

Learning Guide, Lecture Summary, and Worked Examples

Aims

The aims of this course are to investigate how biological nervous systems accomplish the goals of machine intelligence but while using radically different strategies, architectures, and hardware; and to investigate how artificial neural systems can be designed that try to emulate some of those biological principles in the hope of capturing some of their performance.

Lectures

- **Natural versus artificial substrates of intelligence.** Comparison of the differences between biological and artificial intelligence in terms of architectures, hardware, and strategies. Levels of analysis; mechanism and explanation; philosophical issues. Basic neural network architectures compared with rule-based or symbolic approaches to learning and problem-solving.

- **Neurobiological wetware: architecture and function of the brain.** Human brain architecture. Sensation and perception; learning and memory. What we can learn from neurology of brain trauma; modular organisation and specialisation of function. Aphasias, agnosias, apraxias. How stochastic communications media, unreliable and randomly distributed hardware, slow and asynchronous clocking, and imprecise connectivity blueprints, give us unrivalled performance in real-time tasks involving perception, learning, and motor control.

- **Neural processing and signalling.** Information content of neural signals. Spike generation processes. Neural hardware for both processing and communications. Can the mechanisms for neural processing and signalling be viably separated? Biophysics of nerve cell membranes and differential ionic permeability. Excitable membranes. Logical operators.

- **Stochasticity in neural codes.** Principal Components Analysis of spike trains. Evidence for detailed temporal modulation as a neural coding and communications strategy. Is stochasticity also a fundamental neural computing strategy for searching large solution spaces, entertaining candidate hypotheses about patterns, and memory retrieval? John von Neumann's conjecture. Simulated annealing.

- **Neural operators that encode, analyse, and represent image structure.** How the mammalian visual system, from retina to brain, extracts information from optical images and sequences of them to make sense of the world. Description and modelling of neural operators in engineering terms as filters, coders, compressors, and pattern matchers.

- **Cognition and evolution. Neuropsychology of face recognition.** The sorts of tasks, primarily social, that shaped the evolution of human brains. The computational load of social cognition as the driving factor for the evolution of large brains. How the degrees-of-freedom within faces and between faces are extracted and encoded by specialised areas of the brain concerned with the detection, recognition, and interpretation of faces and facial expressions. Efforts to simulate these faculties in artificial systems.

- **Artificial neural networks for pattern recognition.** A brief history of artificial neural networks and some successful applications. Central concepts of learning from data, and foundations in probability theory. Regression and classification problems viewed as non-linear mappings. Analogy with polynomial curve fitting. General "linear" models. The curse of dimensionality, and the need for adaptive basis functions.
- **Probabilistic inference.** Bayesian and frequentist views of probability and uncertainty. Regression and classification expressed in terms of probability distributions. Density estimation. Likelihood function and maximum likelihood. Neural network output viewed as conditional mean.
- **Network models for classification and decision theory.** Probabilistic formulation of classification problems. Prior and posterior probabilities. Decision theory and minimum misclassification rate. The distinction between inference and decision. Estimation of posterior probabilities compared with the use of discriminant functions. Neural networks as estimators of posterior probabilities.

Objectives

At the end of the course students should:

- be able to describe key aspects of brain function and neural processing in terms of computation, architecture, and communication.
- be able to analyse the viability of distinctions such as computing vs communicating, signal vs noise, and algorithm vs hardware, when these dichotomies from Computer Science are applied to the brain.
- understand the neurobiological mechanisms of vision well enough to think of ways to implement them in machine vision.
- understand basic principles of the design and function of artificial neural networks that learn from examples and solve problems in classification and pattern recognition.

Reference books

Aleksander, I. (1989). *Neural Computing Architectures.* North Oxford Academic Press.

Bishop, C.M. (1995). *Neural Networks for Pattern Recognition.* Oxford University Press.

Haykin, S. (1994). *Neural Networks: A Comprehensive Foundation.* Macmillan.

Hecht-Nielsen, R. (1991). *Neurocomputing.* Addison-Wesley.

Exercise 1

List five critical respects in which the operating principles that are apparent in biological nervous tissue differ from those that apply in current computers, and in each case comment upon how these might explain key differences in performance such as adaptability, speed, fault-tolerance, and ability to solve ill-conditioned problems.

Model Answer – Exercise 1

(Five items such as the nine on this list):

1. Neural tissue is asynchronous, and there is no master clock; time is an analog (continuous) variable in neural computing. Computers are synchronous and everything happens on the edges of discrete clock ticks.
2. Neural tissue involves random connectivity, with no master blueprint; connectivity in computers is precisely specified.
3. No single element is irreplaceable in neural tissue, and function appears unaffected by the deaths of 10,000s of neurones every day. Not so in computing machines.
4. Neural processing is highly distributed, seeming to show equipotentiality and recruitment of neural machinery as needed. Less true in computers, which have rigid hardware functional specification (e.g. memory vs. ALU).
5. The elementary 'cycle time' of neurones (i.e. their time constant) is of the order of one millisecond, whereas silicon gate times are on the order of one nanosecond, i.e., a million times faster.
6. The number of elementary functional units in brains (10^{11} neurones, 10^{15} synapses) exceeds by a factor of more than 1,000 those in computers.
7. Brains are able to tolerate ambiguity, to learn on the basis of very impoverished and disorganized data (e.g. to learn a grammar from random samples of poorly structured, unsystematic, natural language); whereas much more precisely formatted and structured data and rules are required by computers.
8. Communications media in neural tissue involve stochastic codes; those in computing machines are deterministic and formally structured.
9. Brains do not seem to have formal symbolic or numerical representations, unlike computers (although humans certainly can perform symbolic manipulation, if only very slowly by comparison). In general, it appears that those tasks for which we humans have *cognitive penetrance* (i.e. an understanding of how we do them, like mental arithmetic), we are not very efficient at, in comparison to machines; but those tasks at which we excel, and machines can hardly perform at all (e.g. adaptive behaviour in unpredictable or novel environments, or face recognition, or language acquisition), are tasks for which we have virtually no cognitive penetrance (ability to explain how we do them).

Exercise 2

1. Illustrate how stochasticity can be used in artificial neural networks to solve, at least in an asymptotic sense, problems that would otherwise be intractable. Name at least two such stochastic engines; describe the role of stochasticity in each; and identify the kinds of problems that such artificial neural devices seem able to solve.

2. Illustrate the evidence for stochasticity in natural nervous systems, and comment on the role that it might play in neurobiological function. What is the case supporting John von Neumann's deathbed prediction that it might be a computational engine for the nervous system, rather than just random noise? Describe at least one experiment involving neural tissue in support of this theory.

Model Answer – Exercise 2

1. Stochasticity offers an opportunity to explore very large problem spaces in search of a solution or a globally optimal match by *blind variation and selective retention.* Random variation is analogous to "temperature," or perturbations in state whose average variance specifies a relationship between entropy and energy that is analogous to temperature. The use of random variation ensures that (in a statistical sense) all corners of the state space can be represented and/or explored. This is an approach to solving NP-complete problems that relies upon asymptotic convergence of expected values, rather than deterministic convergence of an algorithm upon the solution. Its prime disadvantages are (i) very slow operation; and (ii) no guarantee of finding the optimal solution.

Two examples of stochastic engines: Simulated Annealing; and Genetic Algorithms. Role of stochasticity in SA: temperature that declines according to a specific annealing schedule, representing random jumps through state space but with declining average amplitude, so that improvements are more likely, but traps are avoided in the long-term. Role of stochasticity in GA's: mutations of the genotype, with those that increase fitness being retained. Type of problem approached with SA: the Travelling Salesman Problem. Type of problem approached with GA: Monte Carlo combinatorial optimization.

2. Sequences of nerve action potentials are stochastic time-series whose random structure resembles, to first order, a variable-rate Poisson process. The inter-arrival time distributions tend to be exponentials, and the counting distributions tend to be gamma distributions.

von Neumann's prediction that stochasiticity may play an important role in neurobiological function is supported by the fact that seemingly *identical* visual stimuli can generate very different spike sequences from the same neurone in successive presentations, as though possibly different hypotheses were being "entertained" about the pattern and compared with the responses from other neurones. A second argument is that if noise were disadvantageous, then it should quickly have been eliminated in evolution (i.e. Nature could easily have evolved less "noisy" membranes than those with the electrophysiological properties of nerve cells). One set of experiments supporting the hypothesis are those of Optican and Richmond, in which a Principle Components Analysis (PCA) of the response sequences of neurones in the infero-temporal (IT) lobe of macaque monkeys were recorded while they looked at various orthogonal visual patterns (2D Walsh functions). It was found that there are systematic eigenfunctions (shared among large populations of IT neurones) of spike train variation that seem to form temporal-modulation codes for spatial patterns. The conclusion was that much more than just the "mean firing rate" of neurones matters, and that these higher moments of (otherwise seemingly random variation in firing) were in fact responsible for about two-thirds of all the information being encoded.

Exercise 3

Discuss how neural operators that encode, analyze, and represent image structure in natural visual systems can be implemented in artificial neural networks. Include the following issues:

- receptive field structure
- adaptiveness
- perceptual learning
- hierarchies of tuning variables in successive layers
- the introduction of new signal processing dimensions and of non-linearities in successive layers
- wavelet codes for extracting pattern information in highly compressed form
- self-similarity of weighting functions
- associative memory or content-addressable memory for recognizing patterns such as faces and eliciting appropriate response sequences

Model Answer – Exercise 3

- Receptive Field Concept: a linear combination of image pixels is taken by some neurone, with weights (either positive or negative) to produce a sum which determines the output response of the neurone. The Receptive Field constitutes that region of visual space in which information can directly influence the neurone. Its distribution of weights determines primarily the functionality of the neurone. These elements of natural nervous systems are the standard elements of Artificial Neural Networks (ANN's).

- Adaptiveness: the summation weights over the receptive field can be adaptive, controlled by higher-order neural processes, which may be hormonal or neuro-peptides in the case of natural nervous systems. In ANN's, the standard model for adaptiveness is the ADELINE (Adaptive Linear Combiner), and involves global feedback control over all gain parameters in the network.

- Perceptual learning involves the modification of synaptic strengths or other gain factors in response to visual experience, such as the learning of a particular face. In natural visual systems, the almost real-time modification of receptive field properties of neurones has been observed, depending upon other stimulation occuring in (possibly remote) parts of visual space.

- Hierarchies of tuning variables: In the retina and the lateral geniculate nucleus, the spatial tuning variables for visual neurones are primarily size and center-surround structure. But in the visual cortex, the new tuning variable of orientation selectivity is introduced. Another one is stereoscopic selectivity (disparity tuning). At still higher levels in the infero-temporal cortex, still more abstract selectivities emerge, such as neurones tuned to detect faces and to be responsive even to particular aspects of facial expression, such as the gaze.

- Beyond the primary visual cortex, all neurones have primarily non-linear response selectivities. But up to the level of "simple cells" in V1, many response properties can be described as linear, or quasi-linear.

- Cortical receptive field structure of simple cells can be described by a family of 2D wavelets, which have five primary degrees-of-freedom: (1) and (2) the X,Y coordinates of the neurone's receptive field in visual space; (3) the size of its receptive field; (4) its orientation of modulation between excitatory and inhibitory regions; and (5) its phase, or symmetry. These wavelet properties generate complete representations for image structure, and moreover do so in highly compressed form because the wavelets serve as decorrelators.

- To a good approximation, the receptive field profiles of different visual neurones in this family are *self-similar* (related to each other by dilation, rotation, and translation).

- Many neurones show the property of associative recall (or content addressability), in the sense that even just very partial information such as a small portion of an occluded face seems able to suffice to generate the full response of that neurone when presented with the entire face. This idea has been exploited in "Hopfield networks" of artificial neurones for fault-tolerant and content-addressable visual processing and recognition.

Exercise 4

Explain the concepts of the 'curse of dimensionality' and 'intrinsic dimensionality' in the context of pattern recognition. Discuss why models based on linear combinations of fixed basis functions of the form

$$y(\mathbf{x}) = \sum_j w_j \phi_j(\mathbf{x})$$

suffer from the curse of dimensionality, and explain how neural networks, which use adaptive basis functions, overcome this problem.

Exercise 5

By using the example of polynomial curve fitting through noisy data, explain the concept of generalization. You should include a discussion of the role of model complexity, an explanation of why there is an optimum level of complexity for a given data set, and a discussion of how you would expect this optimum complexity to depend on the size of the data set.

Exercise 6

Consider a regularized error function of the form

$$\widetilde{E}(\mathbf{w}) = E(\mathbf{w}) + \nu\Omega(\mathbf{w})$$

and suppose that the unregularized error E is minimized by a weight vector $\mathbf{w}^*$. Show that, if the regularization coefficient ν is small, the weight vector $\widetilde{\mathbf{w}}$ which minimizes the regularized error satisfies

$$\widetilde{\mathbf{w}} \simeq \mathbf{w}^* - \nu\mathbf{H}^{-1}\nabla\Omega$$

where the gradient $\nabla\Omega$ and the Hessian $\mathbf{H} = \nabla\nabla E$ are evaluated at $\mathbf{w} = \mathbf{w}^*$.

Exercise 7

Suppose we have three boxes each containing a mixture of apples, oranges and pears. Box 1 contains 8 apples, 1 orange and 1 pear, box 2 contains 2 apples, 8 oranges and no pears, and box 3 contains 2 apples, 4 oranges and 14 pears. A box is chosen at random with probabilities 0.1 for box 1, 0.3 for box 2 and 0.6 for box 3, and an item is withdrawn at random from the selected box. Find the probability that the item will be an apple. If the item is indeed found to be an apple, evaluate the probability that the box chosen was box 1.

Model Answer – Exercise 4

The term *curse of dimensionality* refers to a range of phenomena whereby certain pattern recognition techniques require quantities of training data which increase exponentially with the number of input variables. Regression models consisting of linear combinations of fixed basis functions $\phi_i(\mathbf{x})$ are prone to this problem. As a specific example, suppose that each of the input variables is divided into a number of intervals, so that the value of a variable can be specified approximately by saying in which interval it lies. By increasing the number of divisions along each axis we could increase the precision with which the input variables can be described. This leads to a division of the whole input space into a large number of cells. Let us now choose the basis function $\phi_j(\mathbf{x})$ to be zero everywhere except within the jth cell (over which it is assumed to be constant). Suppose we are given data points in each cell, together with corresponding values for the output variable. If we are given a new point in the input space, we can determine a corresponding value for y by finding which cell the point falls in, and then returning the average value of y for all of the training points which lie in that cell. We see that, if each input variable is divided into M divisions, the total number of cells is M^d and this grows *exponentially* with the dimensionality of the input space. Since each cell must contain at least one data point, this implies that the quantity of training data needed to specify the mapping also grows exponentially. Although the situation can be improved somewhat by better choices for the basis functions, the underlying difficulties remain as long as the basis functions are chosen independently of the problem being solved. For most real problems, however, the input variables will have significant (often non-linear) correlations, so that the data does not fill the input space uniformly, but rather is confined (approximately) to a lower-dimensional manifold whose dimensionality is called the *intrinsic dimensionality* of the data set. Furthermore, the output value may have significant dependence only on particular directions within this manifold. If the basis functions depend on adjustable parameters, they can adapt to the position and shape of the manifold and to the dependence of the output variable(s) on the inputs. The number of such basis functions required to learn a suitable input-output function will depend primarily on the complexity of the non-linear mapping, and not simply on the dimensionality of the input space.

Model Answer – Exercise 5

The goal in solving a pattern recognition problem is to achieve accurate predictions for new data. This is known as *generalization*, and it is important to understand that good performance on the training does not necessarily imply good generalization (although poor performance on the training data will almost certainly result in equally poor generalization). A practical method of assessing generalization is to partition the available data into a training set and a separate validation set. The training set is used to optimize the parameters of the model, while the validation set is used to assess generalization performance. An important factor governing generalization is the complexity (or flexibility) of the model. In the case of a polynomial, the complexity is governed by the order of the polynomial as this controls the number of adaptive parameters (corresponding to the coefficients in the polynomial). Polynomials of order M include polynomials of all orders $M' < M$ as special cases (obtained by setting the corresponding coefficients to zero), so an increase in the order of the polynomial will never result in an increase in training set error, and will often result in a decrease. Naively we might expect that the same thing would hold true also for the error measured on the validation set. However, in practice this is not the case since we must deal with a data set of finite size in which the data values are noisy. Suppose we use a polynomial to fit a data set in which the output variable has a roughly quadratic dependence on the input variable, and where the data values are corrupted by noise. A linear (first order) polynomial will give a poor fit to the training data, and will also have poor generalization, since it is unable to capture the non-linearity in the input-output mapping. A quadratic (second-order) polynomial will give a smaller error on both training and validation sets. Polynomials of higher order will give still smaller training set error, and will even give zero error on the training set if the number of coefficients equals the number of training data points. However, they do so at the expense of fitting the noise on the data as well as its underlying trend. Since the noise component of the validation data is independent of that on the training data, the result is an increase in the validation set error. Figure 1 shows a schematic illustration of the error of a trained model measured with respect to the training set and also with respect to an independent validation set.

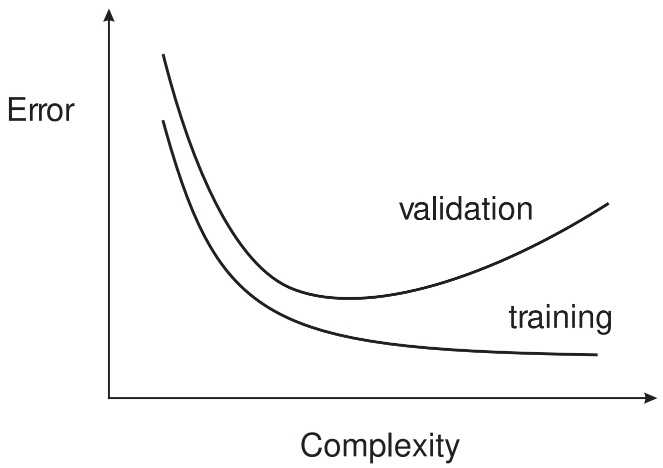

Figure 1: Schematic illustration of the behaviour of training set and validation set error versus model complexity.

Model Answer – Exercise 6

Taking the gradient of the regularized error function and evaluating it at $\mathbf{w} = \tilde{\mathbf{w}}$ we obtain

$$\nabla \widetilde{E}(\tilde{\mathbf{w}}) = \nabla E(\tilde{\mathbf{w}}) + \nu \nabla \Omega(\tilde{\mathbf{w}}) = 0$$

where we have made use of the fact that $\nabla \widetilde{E}(\tilde{\mathbf{w}}) = 0$ since $\tilde{\mathbf{w}}$ is a stationary point of $\widetilde{E}$. Now we Taylor expand ∇E and $\nabla \Omega$ around $\mathbf{w}^*$ to give

$$0 = \nabla E(\mathbf{w}^*) + \nabla\nabla E(\mathbf{w}^*)(\tilde{\mathbf{w}} - \mathbf{w}^*) + \nu \nabla \Omega(\mathbf{w}^*) + O(\nu^2).$$

Using $\nabla E(\mathbf{w}^*) = 0$, and solving for $\tilde{\mathbf{w}}$, we then obtain

$$\tilde{\mathbf{w}} = \mathbf{w}^* - \nu \mathbf{H}^{-1} \nabla \Omega(\mathbf{w}^*)$$

where $\mathbf{H} = \nabla\nabla E(\mathbf{w}^*)$.

Model Answer – Exercise 7

The prior probabilities for the three boxes are $P(1) = 0.1$, $P(2) = 0.3$ and $P(3) = 0.6$. Similarly the conditional probabilities for apple given each of the three boxes are $P(A|1) = 0.8$, $P(A|2) = 0.2$ and $P(A|3) = 0.1$. From these we can first evaluate the unconditional probability of choosing an apple by marginalizing over the choice of box

$$\begin{aligned} P(A) &= P(A|1)P(1) + P(A|2)P(2) + P(A|3)P(3) \\ &= 0.08 + 0.06 + 0.06 = 0.2 \end{aligned}$$

The posterior probability of having chosen box 1 is then obtained from Bayes' theorem

$$P(1|A) = \frac{P(A|1)P(1)}{P(A)} = \frac{0.8 \times 0.1}{0.2} = 0.4$$

Exercise 8

Consider a network having a vector $\mathbf{x}$ of inputs and a single output $y(\mathbf{x})$ in which the output value represents the posterior probability of class membership for a binary classification problem. The error function for such a network can be written

$$E = -\sum_{n=1}^{N} \{t_n \ln y(\mathbf{x}_n) + (1 - t_n) \ln(1 - y(\mathbf{x}_n))\}$$

where $t_n \in \{0, 1\}$ is the target value corresponding to input pattern $\mathbf{x}_n$, and N is the total number of patterns. In the limit $N \to \infty$, the average error per data point takes the form

$$E = -\iint \{t \ln y(\mathbf{x}) + (1 - t) \ln(1 - y(\mathbf{x}))\}\, p(t|\mathbf{x}) p(\mathbf{x})\, dt\, d\mathbf{x}. \tag{1}$$

By functional differentiation of (1), show that, for a network model with unlimited flexibility, the minimum of E occurs when

$$y(\mathbf{x}) = \int t p(t|\mathbf{x})\, dt$$

so that the network output represents the conditional average of the target data. Next consider a network with multiple outputs representing posterior probabilities for several mutually exclusive classes, for which the error function is given by

$$E = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_k(\mathbf{x}_n)$$

where the target values have a 1–of–K coding scheme so that $t_{nk} = \delta_{kl}$ for a pattern n from class l. (Here $\delta_{kl} = 1$ if $k = l$ and $\delta_{kl} = 0$ otherwise.) Write down the average error per data point in the limit $N \to \infty$, and by functional differentiation show that the minimum occurs when the network outputs are again given by the conditional averages of the corresponding target variables. Hint: remember that the network outputs are constrained to sum to unity, so express the outputs in terms of the softmax activation function

$$y_k = \frac{\exp(a_k)}{\sum_l \exp(a_l)}$$

and perform the differentiation with respect to the $\{a_k(\mathbf{x})\}$.

Model Answer – Exercise 8

If we start from the error function in the form

$$E = -\iint \{t \ln y(\mathbf{x}) + (1-t)\ln(1-y(\mathbf{x}))\}\, p(t|\mathbf{x})p(\mathbf{x})\, dt\, d\mathbf{x}$$

and set the functional derivative with respect to $y(\mathbf{x})$ equal to zero we obtain

$$\frac{\delta E}{\delta y(\mathbf{x})} = -\int \left\{ \frac{t - y(\mathbf{x})}{y(\mathbf{x})(1-y(\mathbf{x}))} \right\} p(t|\mathbf{x})p(\mathbf{x})\, dt.$$

Provided $p(\mathbf{x}) \neq 0$ we can solve for $y(\mathbf{x})$ to give

$$y(\mathbf{x}) = \int t p(t|\mathbf{x})\, dt$$

where we have used the normalization property $\int p(t|\mathbf{x})\, dt = 1$ for the conditional distribution. For the case of multiple classes, we can again take the limit $N \to \infty$, so that the error function per data point becomes

$$E = -\iint t \ln\{y(\mathbf{x})\} p(t|\mathbf{x})p(\mathbf{x})\, dt\, d\mathbf{x}.$$

To evaluate the functional derivatives with respect to $a_k(\mathbf{x})$ we first note that, for the softmax activation function

$$\frac{\partial y_k}{\partial a_m} = \frac{\partial}{\partial a_m} \left\{ \frac{\exp(a_k)}{\sum_l \exp(a_l)} \right\} = y_k \delta_{km} - y_k y_m$$

where $\delta_{km} = 1$ if $k = m$ and $\delta_{km} = 0$ otherwise. Hence we obtain

$$\begin{aligned}
\frac{\delta E}{\delta a_m(\mathbf{x})} &= \sum_{k=1}^{K} \int \frac{\delta E}{\delta y_k(\mathbf{x}')} \frac{\delta y_k(\mathbf{x}')}{\delta a_m(\mathbf{x})}\, d\mathbf{x}' \\
&= -\int \{t_k - y_k(\mathbf{x})\}\, p(t_k|\mathbf{x})p(\mathbf{x})\, dt_k.
\end{aligned}$$

Assuming $p(\mathbf{x}) \neq 0$ we can solve for $y_k(\mathbf{x})$ to obtain

$$y_k(\mathbf{x}) = \int t_k p(t_k|\mathbf{x})\, dt_k$$

which again is the conditional average of the target data, conditioned on the input vector.

Exercise 9

Consider a cancer screening application, based on medical images, in which only 1 person in 1000 in the population to be screened has cancer. Suppose that a neural network has been trained on a data set consisting of equal numbers of 'cancer' and 'normal' images, and that the outputs of the network represent the corresponding posterior probabilities $P(C|\mathbf{x})$ and $P(N|\mathbf{x})$ where $C \equiv$ 'cancer' and $N \equiv$ 'normal'. Assume that the loss in classifying 'cancer' as 'normal' is 500 times larger than the loss in classifying 'normal' as 'cancer' (with no loss for correct decisions). Explain clearly how you would use the outputs of the network to assign a new image to one of the classes so as to minimize the expected (i.e. average) loss. If you were also permitted to reject some fraction of the images, explain what the reject criterion would be in order again to minimize the expected loss.

Model Answer – Exercise 9

The prior probabilities of cancer and normal are $P(C) = 10^{-3}$ and $P(N) = 1 - 10^{-3}$ respectively. From Bayes' theorem we know that the posterior probabilities $P(C|\mathbf{x})$ and $P(N|\mathbf{x})$ are proportional to the artificial prior probabilities of $\widehat{P}(C) = \widehat{P}(N) = 0.5$ used to train the network. Hence we can find the posterior probabilities $\widetilde{P}(C|\mathbf{x})$ and $\widetilde{P}(N|\mathbf{x})$ corresponding to the true priors by dividing by the old priors and multiplying by the new ones, so that

$$\widetilde{P}(C|\mathbf{x}) \propto P(C|\mathbf{x}) \frac{10^{-3}}{0.5}$$

and similarly for $\widetilde{P}(N|\mathbf{x})$. Normalizing and cancelling the factors of 0.5 we then obtain

$$\widetilde{P}(C|\mathbf{x}) = \frac{P(C|\mathbf{x})}{P(C|\mathbf{x}) + P(N|\mathbf{x})(10^3 - 1)}$$

with an analogous expression for $\widetilde{P}(N|\mathbf{x})$.

We can now use these corrected posterior probabilities to find the decision rule for minimum expected loss. If an input pattern $\mathbf{x}$ is assigned to class C then the expected loss will be

$$\widetilde{P}(N|\mathbf{x})$$

while if the pattern is assigned to class N the expected loss will be

$$500\widetilde{P}(C|\mathbf{x}).$$

Thus to make a minimum expected loss decision we simply evaluate both of these expressions for the new value of $\mathbf{x}$ and assign the corresponding image to the class for which the expected loss is smaller.

Finally, to reject some fraction of the images we choose a threshold value θ and reject an image if the value of the expected loss, when the image is assigned to the class having the smaller loss, is greater than θ. By changing the value of θ we can control the overall fraction of images which will be rejected, so that by increasing θ we reject fewer images, while if $\theta = 0$ all of the images will be rejected.

Exercise 10

In Computer Science, a fundamental distinction has classically been erected between computing and communications. The former creates, requires, or manipulates data, and the latter moves it around. But in living neural systems, this distinction is less easy to establish; a given neurone performs both functions by generating nerve impulses, and it is not clear where to draw the distinction between processing and communication. Still more so with artificial neural networks, where the entire essence of computing is modeled as just changes in connectivity. Flesh out and discuss this issue. Would you argue that some of the limitations of efforts in artificial intelligence have been the result of such a spurious dichotomy?

Model Answer – Exercise 10

Computing by logical and arithmetic operations is based upon deterministic rules which are guaranteed (in a proper algorithm) to lead to a solution by a sequence of state transitions. Apart from moving bits to and from registers or storage locations, the pathways of communications and their properties are classically not part of the analysis. In wet neural systems, there are no (or few) known rules which could be described as "formal," and little or nothing appears to be deterministic. Rather, stochasticity appears to be the best description of membrane properties, signalling events, and overall neural activity. Similarly, the connectivity between and among neurones is not based upon precise blueprints, but rather upon connectivity matrices which are probabilistic both in their wiring and in their connection strengths. It may be that signalling, or communications, among neurones are the essence of wet neural computing, rather than any distinct processing rules which in any way resemble a sequence of instructions. In artificial neural nets, this general view is the basic strategy for learning and problem-solving. Connectivity is everything. Learning occurs by adaptive modification of connection strengths, often following rules which are primarily probabilistic and rarely even remotely formal. The events which underlie neural network computing are analog, or graded, rather than discrete states and transitions among states. Finally, an influential view today is that "physics is computation," meaning that the laws of nature underlying dynamical systems, energy minimization, and entropy flows over time, may be the only way to implement the sorts of computations that are required which cannot readily be reduced to mere symbol manipulation. If the tasks which require solution in artificial intelligence (e.g. vision, or learning) are formally "intractable," as is generally accepted, then this observation may well account for the failure of AI largely to deliver on its promises. Implementing the ill-posed problems of AI instead as optimization problems or as stochastic explorations of huge-dimensional solution spaces may be the key strategy behind wet neural computing, and may indeed be the only way forward for AI.

Exercise 11

Explain the mechanisms and computational significance of nerve impulse generation and transmission. Include the following aspects:

1. Equivalent electrical circuit for nerve cell membrane.
2. How different ion species flow across the membrane, in terms of currents, capacitance, conductances, and voltage-dependence. (Your answer can be qualitative.)
3. Role of positive feedback and voltage-dependent conductances.
4. The respect in which a nerve impulse is a mathematical catastrophe.
5. Approximate time-scale of events, and the speed of nerve impulse propagation.
6. What happens when a propagating nerve impulse reaches an axonal branch.
7. What would happen if two impulses approached each other from opposite directions along a single nerve fibre and collided.
8. How linear operations like integration in space and time can be combined in dendritic trees with logical or Boolean operations such as AND, OR, NOT, and veto.
9. Whether "processing" can be distinguished from "communications," as it is for artificial computing devices.
10. Respects in which stochasticity in nerve impulse time-series may offer computational opportunities that are absent in synchronous deterministic logic.

Model Answer – Exercise 11

1. Equivalent Circuit diagram:

   (see lecture notes)

2. A nerve cell membrane can be modelled in terms of electrical capacitance C and several conductances (or resistances R) specific to particular ions which carry charge across the membrane. These currents I into and out of the nerve cell affect its voltage V in accordance with Ohm's Law for resistance ($I = V/R$) and the law for current flow across a capacitor ($I = C\frac{dV}{dt}$). The charge-carrying ionic species are sodium (Na^+), potassium (K^+), and cloride (Cl^-) ions.

3. The crucial element underlying nerve impulse generation is the fact that the conductances (resistances) for Na^+ and K^+ are not constant, but voltage-dependent. Moreover, these two voltage-dependent conductances have different time courses (time constants). The more the voltage across a nerve cell rises due to Na^+ ions flowing into it, the lower the resistance for Na^+ becomes. (Na^+ current continues to flow until its osmotic concentration gradient is in equilibrium with the voltage gradient.) This positive feedback process causes a voltage spike, which is a nerve impulse.

4. Since the positive feedback process is unstable, causing the voltage to climb higher and higher, faster and faster, it can be described as a catastrophe, rather like an explosion. Once combustion starts in a small corner of a keg of dynamite, matters just get worse and worse. The positive climb of voltage only stops when, on a slower time-scale, K^+ begins to flow in the opposite direction. Once the trans-membrane voltage falls below its threshold, the resting state of ionic concentrations can be restored by ion pumps and the catastrophe (the nerve impulse) is over.

5. The process described above is complete within about 2 milliseconds. There is a refractory period for restoration of ionic equilibrium that lasts for about 1 millisecond, so the fastest frequency of nerve impulses is about 300 Hz. The speed of nerve impulse propagation down an excitable myelinated axon, by saltatory spike propagation, can reach 100 meters per second in warm-blooded vertebrates.

6. A nerve impulse reaching an axonal branch would normally go down both paths, unless vetoed at either one by other shunting synapses. Some axonal branches are "steerable" by remote signalling.

7. The two approaching nerve impulses would annihilate each other when they collided. The minimum refractory period of excitable nerve membrane prevents the impulses from being able to pass through each other, as they would if they were pulses propagating in a linear medium such as air or water or the aether.

8. The linear components of nerve cells (i.e. their capacitance and any non-voltage-dependent resistances) behave as linear integrators, providing linear (but leaky) summation of currents over space and time. However, the fundamentally non-linear interactions at synapses can implement logical operations such as AND, OR, NOT, and veto. The basic factor which

underlies this “logico-linear” combination of signal processing is the mixture of excitable (“logical”) and non-excitable (“linear”) nerve cell membranes.

9. It is very difficult to distinguish between processing and communications in living nervous tissue. The generation and propagation of nerve impulses is the basis for both. A steerable axonal branch particularly illustrates the impossibility of making such a distinction.

10. Stochasticity in nerve impulse time-series may provide a means to search very large spaces for solutions (e.g. to pattern recognition problems) in a way resembling “simulated annealing.” Evolutionary computing (blind variation and selective retention of states) can be the basis of learning in neural networks, and stochasticity may provide the blind variation.

Exercise 12

Explain why probability theory plays a central role in neural computation. Discuss how the problem of classification can be expressed in terms of the estimation of a probability distribution.

Explain what is meant by a *likelihood function* and by the concept of *maximum likelihood.*

Consider a neural network regression model which takes a vector $\mathbf{x}$ of input values and produces a single output $y = f(\mathbf{x}, \mathbf{w})$ where $\mathbf{w}$ denotes the vector of all adjustable parameters ('weights') in the network. Suppose that the conditional probability distribution of the target variable t, given an input vector $\mathbf{x}$, is a Gaussian distribution of the form

$$p(t|\mathbf{x}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\{t - f(\mathbf{x}, \mathbf{w})\}^2}{2\sigma^2}\right)$$

where σ^2 is the variance parameter.

Given a data set of input vectors $\{\mathbf{x_n}\}$, and corresponding target values $\{t_n\}$, where $n = 1, \ldots, N$, write down an expression for the likelihood function, assuming the data points are independent. Hence show that maximization of the likelihood (with respect to $\mathbf{w}$) is equivalent to minimization of a sum-of-squares error function.

Model Answer – Exercise 12

Neural computation deals with problems involving real-world data and must therefore address the issue of *uncertainty*. The uncertainty arises from a variety of sources including noise on the data, mislabelled data, the natural variability of data sources, and overlapping class distributions. Probability theory provides a consistent framework for the quantification of uncertainty, and is unique under a rather general set of axioms.

The goal in classification is to predict the class $\mathcal{C}_k$ of an input pattern, having observed a vector $\mathbf{x}$ of features extracted from that pattern. This can be achieved by estimating the conditional probabilities of each class given the input vector, i.e. $P(\mathcal{C}_k|\mathbf{x})$. The optimal decision rule, in the sense of minimising the average number of mis-classifications, is obtained by assigning each new $\mathbf{x}$ to the class having the largest posterior probability.

The likelihood function, for a particular probabilistic model and a particular observed data set, is defined as the probability of the data set given the model, viewed as a function of the adjustable parameters of the model. Maximum likelihood estimates the parameters to be those values for which the likelihood function is maximized. It therefore gives the parameter values for which the observed data set is the most probable.

Since the data points are assumed to be independent, the likelihood function is given by the product of the densities evaluated for each data point

$$\begin{aligned} \mathcal{L}(\mathbf{w}) &= \prod_{n=1}^{N} p(t_n|\mathbf{x_n}, \mathbf{w}) \\ &= \prod_{n=1}^{N} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{\{t_n - f(\mathbf{x_n}, \mathbf{w})\}^2}{2\sigma^2}\right). \end{aligned}$$

Following the standard convention, we can define an error function by the negative logarithm of the likelihood

$$E(\mathbf{w}) = -\ln \mathcal{L}(\mathbf{w}).$$

Since the negative logarithm is a monotonically decreasing function, maximization of $\mathcal{L}(\mathbf{w})$ is equivalent to minimization of $E(\mathbf{w})$. Hence we obtain

$$E(\mathbf{w}) = \frac{1}{2\sigma^2} \sum_{n=1}^{N} \{t_n - f(\mathbf{x_n}, \mathbf{w})\}^2 + \frac{N}{2} \ln(2\pi\sigma^2)$$

which, up to an additive constant independent of $\mathbf{w}$ and a multiplicative constant also independent of $\mathbf{w}$, is the sum-of-squares error function.

Exercise 13

Give brief explanations of the following terms:

(a) the curse of dimensionality,

(b) the Perceptron,

(c) error back-propagation,

(d) generalisation,

(e) loss matrix.

Model Answer – Exercise 13

(a) *The curse of dimensionality.*

Many simple models used for pattern recognition have the unfortunate property that the number of adaptive parameters in the model grows rapidly, sometimes exponentially, with the number of input variables (i.e. with the dimensionality of the input space). Since the size of the data set must grow with the number of parameters, this leads to the requirement for excessively large data sets, as well as increasing the demands on computational resources. An important class of such models is based on linear combinations of fixed, non-linear basis functions. The worst aspects of the curse of dimensionality in such models can be alleviated, at the expense of greater computational complexity, by allowing the basis functions themselves to be adaptive.

(b) *The Perceptron.*

The Perceptron is a simple neural network model developed in the 1960s by Rosenblatt. He built hardware implementations of the Perceptron, and also proved that the learning algorithm is guaranteed to find an exact solution in a finite number of steps, provided that such a solution exists. The limitations of the Perceptron were studied mathematically by Minsky and Papert.

(c) *Error back-propagation.*

Neural networks consisting of more than one layer of adaptive connections can be trained by error function minimisation using gradient-based optimisation techniques. In order to apply such techniques, it is necessary to evaluate the gradient of the error function with respect to the adaptive parameters in the network. This can be achieved using the chain rule of calculus which leads to the error back-propagation algorithm. The name arises from the graphical interpretation of the algorithm in terms of a backwards propagation of error signals through the network.

(d) *Generalisation.*

The parameters of a neural network model can be determined through minimisation of an appropriate error function. However, the goal of training is not to give good performance on the training data, but instead to give the best performance (in terms of smallest error) on independent, unseen data drawn from the same distribution as the training data. The capability of the model to give good results on unseen data is termed generalisation.

(e) *Loss matrix.*

In many classification problems, different misclassification errors can have different consequences and hence should be assigned different penalties. For example, in a medical screening application the cost of predicting that a patient is normal when in fact they have cancer is much more serious than predicting they have cancer when in fact they are healthy. This effect can be quantified using a loss matrix consisting of penalty values for each possible combination of true class versus predicted class. The elements on the leading diagonal correspond to correct decisions and are usually chosen to be zero. A neural network model can be used to estimate the posterior probabilities of class membership for a given input vector. Simple decision theory then shows that if these posterior probabilities are weighted by the appropriate elements of the loss matrix, then selection of the largest weighted probability represents the optimal classification strategy in the sense of minimising the average loss.

Exercise 14

When using a feed-forward network to solve a classification problem we can interpret the network's outputs as posterior probabilities of class membership, and then subsequently use these probabilities to make classification decisions. Alternatively, we can treat the network as a discriminant function which is used to make the classification decision directly. Discuss the relative merits of these two approaches.

Explain the concept of a likelihood function, and the principle of maximum likelihood.

Consider a feed-forward network which implements a function $y(\mathbf{x}, \mathbf{w})$ in which y is the output variable, $\mathbf{x}$ is the vector of input variables, and $\mathbf{w}$ is the vector of weight parameters. We wish to use this network to solve a classification problem involving two classes $\mathcal{A}$ and $\mathcal{B}$. The value of y, when the network is presented with an input vector $\mathbf{x}$, is to be interpreted as the posterior probability $P(t = 1|\mathbf{x})$ in which $t = 1$ denotes class $\mathcal{A}$ and $t = 0$ denotes class $\mathcal{B}$. Write down the probability distribution of t given y. Use the principle of maximum likelihood to derive an expression for the corresponding error function defined over a set of training data comprising input vectors $\mathbf{x_n}$ and targets t_n, where $n = 1, \ldots, N$.

Write down a suitable form for the output unit activation function $y = g(a)$. Hence evaluate the derivative of $\ln P(t|y)$ with respect to a.

Model Answer – Exercise 14

If a network can provide accurate assessments of posterior probabilities, then these can be used to make optimal decisions and hence will give the same results as the discriminant function. In addition, the posterior probabilities allow the following: (i) corrections can be made for prior probabilities which differ between the data set used to train the network and the test data on which the trained network will be used; (ii) if a non-trivial loss matrix is introduced then optimal decisions can still be made; (iii) a reject option can be applied to improve the misclassification rate, at the expense of not reaching a decision on some fraction of the data. If the prior distributions of the test data change, or if the elements of the loss matrix are changed, then optimal decisions can still be made without the need to re-train the network. The disadvantage of trying to estimate posterior probabilities is that typically much more training data will be required than will be needed to obtain a good discriminant function. If the supply of training data is limited, it is possible for the network to provide an accurate discriminant function, even though its estimates of posterior probabilities may be poor.

The likelihood function (for a parametric model described by a vector of parameters $\mathbf{w}$, together with a given data set D) is given by the probability $p(D|\mathbf{w})$ of the data given $\mathbf{w}$, viewed as a function of $\mathbf{w}$. The principle of maximum likelihood says that $\mathbf{w}$ should be set to those values which maximize the likelihood function; these values correspond to the choice of parameters for which the observed data set is the most probable.

The conditional distribution $p(t|y)$ is given by the Bernoulli distribution

$$P(t|y) = y^t(1-y)^{1-t} \tag{2}$$

and hence satisfies $P(t=1|y) = y$ and $P(t=0|y) = 1-y$. Assuming the data are independent and identically distributed, the likelihood function is given by

$$L = \prod_{n=1}^{N} P(t_n|x_n) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$$

where y_n denotes $y(\mathbf{x}_n, \mathbf{w})$. The corresponding error function $E(\mathbf{w})$ is given by the negative logarithm of the likelihood function, so that

$$E(\mathbf{w}) = -\ln L = -\sum_{n=1}^{N} \{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}.$$

The appropriate choice of output unit activation function is the logistic sigmoid given by

$$y = g(a) = \frac{1}{1+\exp(-a)}. \tag{3}$$

The required derivative is thus given, from (1) and (2), by

$$\frac{\partial}{\partial a} \ln P(t|y) = \frac{\partial \ln P(t|y)}{\partial y}\frac{\partial y}{\partial a} = \frac{y-t}{y(1-y)} \cdot y(1-y) = y - t.$$

Exercise 15

(A)

Explain the key ideas of a Hopfield artificial neural network for content-addressable, associative memory. In explaining how memories are stored and retrieved, be sure to define the notions of:

- configuration space
- connectivity matrix
- stable attractor
- basin of attraction
- network capacity, and its dependence on the number of "neurones"

(B)

1. Marshall as many lines of evidence as you can to support the view that in human vision "what you see is your own 'graphics,' rather than the retinal image as faithfully recorded by photoreceptors in the eye." Explain the significance of this observation for vision theory and for machine vision.

2. Suppose you were trying to design a machine vision system based as closely as possible upon human vision. Would you aim to design in the visual illusions that nearly all people "see" as well? (These include the distortions of geometrical form, angle and relative length illusions, etc.) If such properties emerged as unintended consequences of your vision design, would you consider them to be features, or bugs?

Model Answer – Exercise 15

(A)

A Hopfield content-addressable, associative memory allows patterns to be stored in such a way that each one can be retrieved by using just a portion, or a corrupted version, of it as its "address." The memory functions dynamically by detecting an association between an input pattern and a stored memory, which it then retrieves in its entirety.

The network consists of N "neurones" that are fully connected to each other. Their connections have weights of adjustable strengths, defining an NxN connectivity matrix. This matrix of weights is adjusted during a learning phase that stores the patterns in memory. Each of the N neurones has a binary state, and the N-dimensional space of their states is called the configuration space. A stored "memory" is just a particular point in this N-dimensional configuration space. The training phase consists in setting all the NxN weights such that a particular memory is a stable state consistent with all the other stable states, so that once the network reaches such a state, no neurones will change their state further. Then, whenever a new input pattern first appears (implemented by re-setting the states of the N neurones), the collective state of the network will evolve by each neurone following a threshold rule: The $N-1$ inputs that it receives from all the other neurones (each of whose state is either a $+1$ or a -1), each multiplied by the corresponding weight for that connection, are all added together and compared against a threshold. If this inner product (sum) exceeds the threshold, then the neurone's state is set to $+1$; else it is set to -1. This process reiterates for all the N neurones until a stable state is reached, when no neurone changes its state anymore. This collective state is called an stable attractor. Each stable attractor represents one memory. The range of different input states that will all converge eventually to a particular such stable attractor is called the basin of attraction for that memory. Because the states within this basin of attraction represents similar, but corrupted (or partial) versions of the stable state, a Hopfield memory is capable of "completing" a memory from just parts of it serving as a trigger, rather like happens in human memory. Hence it is deemed content-addressable and associative. It is also capable of overcoming noise and corruption of an input pattern, provided that the input remains within the basin of attraction of the attractor. The network capacity of a Hopfield network is about $(0.15)N$, where N is the number of neurones. This capacity would be greater if only orthogonal patterns were stored.

....Continued....

(B)

1. The purpose of vision is not to reproduce faithfully the 2D retinal image, but rather to construct an internal 3D model of the surrounding world and of the 3D objects that populate it. In this regard, vision is a kind of "inverse graphics." (In graphics, a 3D world model is projected into a 2D screen image; in vision, just the reverse must be accomplished.) Evidence that what we see is our own "graphics:"

- Perceptual size invariance: all the faces in a lecture theatre appear to be roughly the same size. Yet in terms of the 2D retinal image, some faces may actually be 20 or 30 times larger than others, depending on their distance.
- Motion compensation: every eye movement causes the retinal image to shift ballistically, in tremendous sweeps and jerks. Yet the world appears to be stable.
- Homogeneity of spatial resolution: high visual resolution exists only within the fovea (the central 1 or 2 degrees). Yet the world appears to have uniform resolution everywhere.
- Homogeneity of colour signals: the colour-sensitive photoreceptors ("cones") exist only near the fovea; outside the fovea, only black and white information is transduced. Yet the world appears uniformly full of colour.
- Invisibility of the silhouette of the retinal blood vessels: the retinal image is corrupted by a dense tree of opaque blood vessels, lying in front of the actual photoreceptors. Yet we discount their shadows, and "fill-in" the gaps with graphics. Similarly, the blind spot is filled in; we are not aware of this large hole.
- Colour constancy: a banana continues to look yellow, whether reddish light or bluish light is shown on it. We see the spectral reflectance properties of the pigmented surface, rather than the actual wavelength mix received on the retina.
- Interpreted scene properties: we perceptually interpret events like occlusion, or motion and rotation in depth, in terms of solid 3D objects and configurations.

One implication of these observations for machine vision is that we should cast the problems of vision as "inverse problems," for which the goal is to invert the processes of graphics (projections, ray-tracing, radiosity, etc) to try to arrive at the unique 3D world situation that could produce such a 2D signal as the one received on the retina.

2. The geometrical illusions (distortions of form, angle, length, shape) may be inevitable consequences of trying to achieve such goals. They may reveal, for example, the existence of near-range competition and far-range cooperation in the processing of orientation information. Mere fidelity to the retinal image should *not* be a primary goal of vision, as it might be in a sound reproduction system. The goal is to *understand* the world; who cares if this *changes* its appearance (inverting the famous maxim of Karl Marx about changing the world versus understanding it...)

Exercise 16

(A)

1. Define "generalisation" in neural networks that learn from training data, and then are tested on new data. Why should not *all* the available data be used in the training set?

2. Draw a simple connectivity diagram that illustrates the idea of lateral inhibition in a competitive neural network.

3. With another diagram showing plots for input and output, illustrate how lateral inhibition in such a competitive network sharpens any input signal by effectively amplifying its first derivative.

4. What class of multi-layer neural network can be used to overcome the mathematical difficulties caused by intrinsic non-orthogonality of representation in many sensory and control systems?

(B)

The study of neurological trauma to the brain gives clues about its modular organisation and specialisations of function, which may reveal some computational principles.

1. What two fundamental principles of brain function did Karl Lashley's neurological research seem to reveal?

2. What are generally the differences between recovery prospects after a sudden brain trauma (in which all the damage is done at once), versus the same damage done more gradually (e.g. by a growing tumour)?

3. What mechanism might explain this difference?

4. Comment on its possible computational significance in terms of fault-tolerance, circuit adaptability and flexibility.

5. Describe two different types of language-related disorders that can result from trauma to Broca's area or Wernicke's area, and comment on the computational inferences we might draw concerning language processing and linguistic representation.

Model Answer – Exercise 16

(A)

1. The parameters of a neural network model can be set via the minimisation of an appropriate error function. However, the goal of training is not to give good performance on the training data, but instead to give the best performance (in terms of smallest error) on independent, unseen data drawn from the same distribution as the training data. The ability of a neural network to give good results on unseen data is termed generalisation. If all of the available data were used in the training set and the model were made too large in order to fit all of it well, then it may become "over-trained" on this one set and will not generalize well to other sample sets. A simpler model trained just on a subset of the available data may actually show better generalisation.

2. (Diagram page 124)

3. (Diagram page 125)

4. A relaxation network.

(B)

1. The two principles of brain function that emerged from Karl Lashley's neurological research were:

   - The Law of Mass Action – the idea that much of the brain seems to participate in the performance of many tasks.
   - The Law of Equipotentiality – the idea that all brain tissue is basically the same, and has the *potential* to subserve any type of brain task if it is recruited to do so.

2. The brain is far more able to overcome, and to compensate for, an insult that develops or spreads slowly, than one caused by sudden trauma.

3. It is thought that the reason is recruitment: other non-injured tissue that may currently not be utilised is recruited to take over the functions served by the damaged tissue, provided there is time for this process to occur.

4. The implication is that brain circuits can train each other. Brain tissue that is involved in the performance of a specific task can transfer its acquired architecture or configuration to other tissue, during a gradual insult (like a spreading neoplasm). This is a remarkable form of adaptability and circuit flexibility, which constitutes a type of fault-tolerance that is unknown in computer hardware.

5. Any two examples from the following list:

   Broca's aphasia: disorders of speech production. Patients may be unable to read aloud, but can read to themselves and comprehend written material perfectly.

   Wernicke's aphasia: disorders of speech understanding. Such aphasics may be able to read sentences aloud but unable to comprehend the material.

   echolalia: a strong tendency to repeat exactly what is heard.

   conduction phasia: Fluent spontaneous speech, normal comprehension, but inability to repeat what is heard.

   alexia: Lost ability to comprehend written language.

   paralexia: Words are read as other words (substitution).

   dyslexia: Confounded letter combinations when reading.

   Computational significance: such disorders imply a remarkable specialisation of function for the different aspects of language. There is evidence that even the different grammatical parts of speech (such as verbs, nouns, prepositions) are stored in different specific places, because highly localised brain traumas can cause the loss of a particular such grammatical category.

Exercise 17

(A)

Give evidence supporting the view that the main computational load that has shaped the evolution of the human brain is "social computation," with sexual success being the ultimate measure of the value of an algorithm or neural design feature. What implications does this have for:

- The cognitive skills and perceptual faculties that have been selected for in brain evolution, as contrasted with the goals which are the traditional focus of AI.
- The design of face recognition algorithms, which aim to interpret facial expression, gesture, and intent, as well as gender and identity.
- The construction of the theory that other persons have minds, too.
- Models of action, planning, and interaction between self and others.

Comment on whether this "social computation" view of human brain evolution implies that brain science is less relevant to the goals of computer science than is usually thought.

(B)

Answer any 5 of the following 7 short questions:

1. Roughly what is the total number of neurones in the human brain?
2. Roughly what is the total number of synapses in the human brain? How does this compare with the total number of stars in our galaxy, and with the total number of galaxies in the known universe?
3. Why is nerve impulse propagation described as "saltatory," and what purposes are achieved by this method of signalling?
4. What is the approximate speed of nerve impulse propagation in warm-blooded animals, in meters/sec?
5. Why is "white matter" white, what cells are responsible for this, and what purpose do they serve?
6. Name the three principal ions involved in nerve membrane current flows, and identify which two of them transit through voltage-dependent conductances.
7. What causes the refractory deadtime of about 1 msec after each nerve impulse?

Model Answer – Exercise 17

(A)
The neural faculties that have been selected for in human brain evolution are those which optimised sexual success and thereby gene transmission; that is how evolution works. These skills focus on: being charming and seductive; outwitting one's rivals; deceiving others; being able to figure out what others are thinking; out-guessing their intentions and contradicting their plans; manipulating others' desires; forming and maintaining alliances; in short, the whole set of social skills cultivated and achieved by ambitious and upwardly mobile humans. (The time scale of evolution is such that no significant brain evolution can have occured in the short recent period since the dawn of civilisation.)

Superior execution of the above social tasks leads to higher status in a power hierarchy in social primates, and therefore to better mating privileges. When the ratio of brain size to body size is compared among various primate species, it is most highly correlated with the characteristic size of social groups for that species. (For purposes of this issue, it does not matter which factor is regarded as 'cause' and which as 'effect.') From smallest to largest, some typical examples of primate social group sizes are: orangutans (3); baboons (20); and chimpanzees (100). It appears that the need to compute and monitor social factors, keeping track of one's con-specifics and of one's own place among them, monitoring others' gaze and actions and modelling their intentions, were important factors determining relative brain size and brain evolution.

Other specific forms of evidence include: the existence of brain areas specialized for processing facial information, including expression, gesture, and intent, as well as facial identity; neurones discovered that are sensitive to the gaze direction of other primates, and especially to eye contact; and evoked brain potentials (electroencephalograms) showing signals sensitive to social power hierarchy (relative position of self and other) in both human and non-human primates.

If such forms of "social computation" are indeed the major computational load of the human brain,

- The traditional goals of AI have not been formulated with a corresponding focus.
- Face processing algorithms, if aimed at simulating human performance with faces, should focus much more on reading social intent and expression than mere identity.
- A crucial step was the evolution of the perspective that other people have minds, too. This (and indeed consciousness itself) would not have been necessary in evolution were it not for the demands of social computation.
- The notion of agency (that others are agents, with their own plans and intentions), and that two agents might need to interact, either in conflict or in cooperation, is perhaps the computational origin of human social psychology.

These perspectives highlight a divergence between the computational goals that have shaped the brain, and those of machine computation. Thus they may limit the current mutual relevance between these two subjects of study. If we wish to "make machines more like humans, rather than humans more like machines," then some consideration and reformulation of goals along these lines may be necessary.

(B)

1. There are roughly 10^{11} neurones in the human brain.
2. There are roughly 10^{15} synapses in the human brain, i.e. about 10,000 times more than the total number of stars in our galaxy, and about 10,000 times more than the total number of galaxies in the known universe.
3. Nerve impulse propagation is described as "saltatory" because when propagating down a single axon, it "jumps" between Nodes of Ranvier which are the isolated patches of excitable membrane separated by about 0.1 mm. The axonal membrane between the Nodes of Ranvier is insulated, and contains no excitable patches. The purposes served by this arrangement are (i) greater speed of impulse propagation, and (ii) less energy consumption by the process, since the dissipative trans-membrane current flows are limited to discrete patches.
4. The approximate speed of nerve impulse propagation is 100 meters/sec in warm-blooded animals.
5. "White matter" consists of the myelinated axons which are the communications channels between neurones. It is white because of myelin insulation, which prevents current leakage and short-circuits. Each axonal fibre is wrapped around many times by insulating Schwann cells.
6. The three principal ions involved in nerve membrane current flows are sodium (Na^+), potassium (K^+), and cloride (Cl^-) ions. Of these three, Na^+ and K^+ transit through voltage-dependent conductances.
7. The time required to restore ionic equilibrium to the resting state concentrations of these three ion species is about 1 msec. Therefore most neurones cannot generate nerve impulses at frequencies exceeding 1 KHz.