Using Neural Networks to Model Conditional Multivariate Densities

Peter M Williams
School of Cognitive and Computing Sciences
University of Sussex
Falmer, Brighton BN1 9QH
email: peterw@cogs.susx.ac.uk

CSRP 371

February 15, 1995

Abstract

Neural network outputs are interpreted as parameters of statistical distributions. This allows us to fit conditional distributions in which the parameters depend on the inputs to the network. We exploit this in modelling multivariate data, including the univariate case, in which there may be input-dependent (e.g. time-dependent) correlations between output components. This provides a novel way of modelling conditional correlation as well as providing input-dependent (local) error bars.

1 Introduction

Neural networks provide a way of modelling the statistical relationship between a dependent variable Y and an independent variable X. For example, X could be financial data up to a certain time and Y could be a future stock index, exchange rate, option price etc. Alternatively X could represent geophysical features of a prospect and Y could represent mineralization at a certain depth. In general X and Y can be vectors of continuous or discrete quantities.

Suppose that the conditional distribution of Y belongs to a family of distributions characterised by a finite set of parameters which are functions of conditioning values of X. These functions, which in general will be non-linear, can then be modelled by a neural network. For discrete distributions this approach has been known for some time in the form of the softmax rule (Bridle, 1990). Bishop (1994) extends this framework to absolutely continuous distributions, in particular to the case of finite Gaussian mixtures. The case of a single kernel is treated independently by Nix and Weigend (1995). Bishop uses radial kernels though it is straightforward to extend the approach to Gaussians with diagonal covariance matrices. The purpose

of this paper is to consider the case of multivariate data in which the conditional covariance matrix may be non-diagonal.

2 Multivariate data

The conditional distribution of the n-dimensional quantity Y given $X = x$ is assumed to be described by the multivariate Gaussian density

$$P(y \mid x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp\left\{-\tfrac{1}{2}(y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)\right\} \qquad (1)$$

where $\mu(x)$ is the vector of conditional means and $\Sigma(x)$ is the conditional covariance matrix. Both μ and Σ are understood to be functions of x in a way that depends on the outputs of a neural network when the conditioning vector x is given as input.

It is assumed that the network has linear output units and that μ and Σ are determined by the activations of these units. We now discuss the link between network outputs and the components of μ and Σ. The **mean** presents no problem. The network will be required to have n output units whose activations, $\{z_i^\mu\}$ say,

To represent the matrix A we stipulate that the network is provided with an additional set of *dispersion* output units whose activations $\{z_i^{\pi}\}$ and $\{z_{ij}^{\alpha}\}$ are related to the elements of A by

$$\alpha_{ii} = \exp$$

and partial derivatives with respect to

4.1 Univariate data

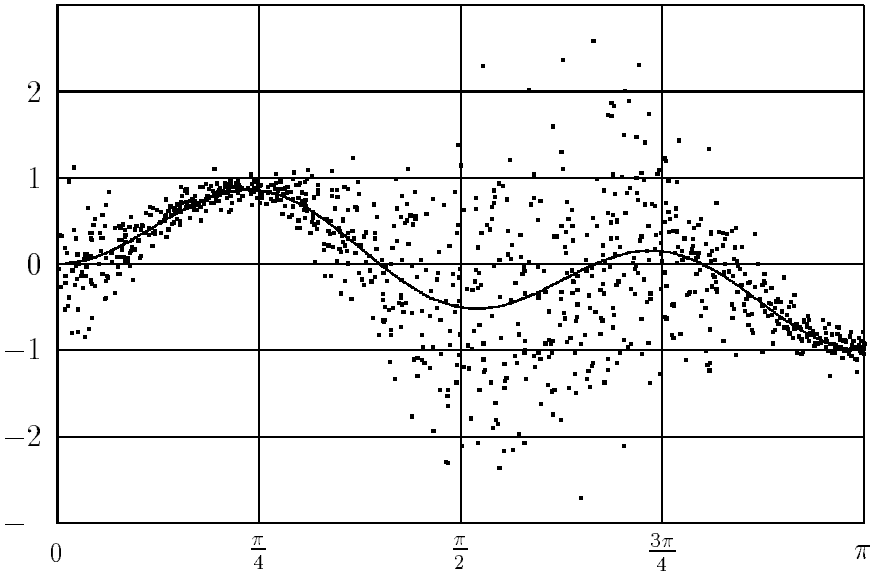
Weigend and Nix (1994) discuss univariate data ($n = 1$) drawn from normal distributions $N(\mu, \sigma)$ with means

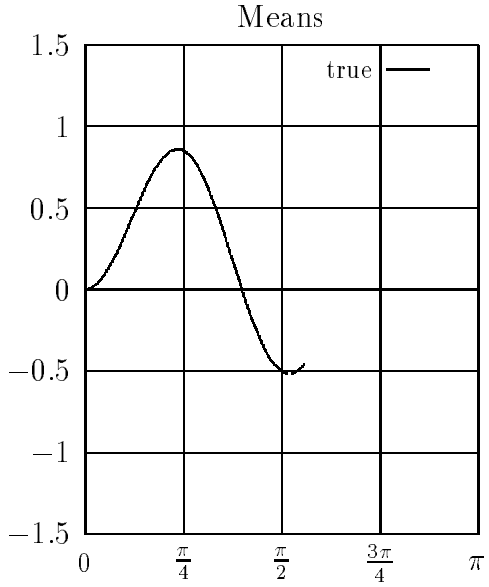
$$\mu(x) = \sin(2.5\,x)\sin(1.5\,x)$$

and variances

$$\sigma^2(x) = 0.01 + 0.25\,[1 - \sin(2.5\,x)]^2\,.$$

1000 training examples were generated using this example with x drawn randomly from a uniform distribution on $[0, \pi]$. The training set is shown in Figure 1. Results are shown in Figure 2. These were obtained using a simple fully connected - layer network with 1 input unit, 10 hidden units and 2 output units. Networks were trained using the optimisation and regularisation algorithms of Williams (1991, 1995) which pruned the


2
1
0
−1
−2
−
0
$\frac{\pi}{4}$
$\frac{\pi}{2}$
$\frac{3\pi}{4}$
π


Means
1.5
1
0.5
0
−0.5
−1
−1.5
true
0
$\frac{\pi}{4}$
$\frac{\pi}{2}$
$\frac{3\pi}{4}$
π

5 Conclusion

Modelling correlation inevitably requires larger

Neal, R. M. 1995. *Bayesian Learning for Neural Networks.* Ph.D. thesis, Graduate Department of Computer Science, University of Toronto.

Nix, D. A., and Weigend, A. S. 1995. Local error bars for nonlinear regression and time series prediction. In Tesauro, G., Touretzky, D. S., and Leen, T. K., eds., *Advances in Neural Information Processing Systems 7.* MIT Press.

Press, W. H., Flannery, B. P., Teukolsky, S. A., and Vetterling, W. T. 1992. *Numerical Recipes in C* (2nd edition). Cambridge University Press.

Rao, C. R. 197 . *Linear Statistical Inference and Its Applications* (2nd edition). Wiley.

Strang, G. 1988. *Linear Algebra and its Applications* (rd edition). Harcourt Brace Jovanovich.

Weigend, A. S., and Nix, D. A. 1994. Predictions with confidence intervals (local error bars). In *Proceedings of the International Conference on Neural Information Processing*, pp. 847–852 Seoul, Korea.

Williams, P. M. 1991. A Marquardt algorithm for choosing the step-size in backpropagation learning with conjugate gradients. Cognitive Science Research Paper CSRP 229, University of Sussex.

Williams, P. M. 1995. Bayesian regularization and pruning using a Laplace prior. *Neural Computation*, 7, 117–14 .