Error Functions for Prediction of Episodes of Poor Air Quality

Robert J. Foxall*, Gavin C. Cawley*, Stephen R. Dorling†
and Danilo P. Mandic‡

*School of Information Systems
†School of Environmental Sciences
University of East Anglia
Norwich NR4 7TJ, U.K.
{rjf,gcc}@sys.uea.ac.uk

‡Dept Electrical and Electronic Engineering
Imperial College of Science,
Technology and Medicine
London SW7 2BT, U.K.
d.mandic@ic.ac.uk

Abstract. Prediction of episodes of poor air quality using artificial neural networks is investigated. Logistic regression, conventional sum-of-squares regression and heteroscedastic sum-of-squares regression are employed for the task of predicting real-life episodes of poor air quality in urban Belfast due to SO_2. In each case, a Bayesian regularisation scheme is used to prevent over-fitting of the training data and to provide pruning of redundant model parameters. Non-linear models assuming a heteroscedastic Gaussian noise process are shown to provide the best predictors of pollutant concentration of the methods investigated.

1 Introduction

Belfast is unusual within the U.K. in that a significant fraction of the city's domestic heating is derived from coal burning, resulting from limited availability of natural gas. This leads to relatively low-level SO_2 emission, the efficient dispersion of which is highly dependent on meteorological conditions. Episodes of high ground-level SO_2 concentrations caused by emissions from tall stacks are mostly short-lived, however the longevity of SO_2 episodes caused by low-level emission may be more extended; meteorological conditions which are unconducive to efficient dispersion may persist for a period of hours to days. Given the health implications of exposure to high concentrations of SO_2, it is important to develop accurate forecast models for both the occurrence and severity of episodes of poor air quality. An ideal model should therefore produce and accurate forecast of the expected concentration of a given pollutant *and* some means of estimating the probability that the observed concentration will exceed a preset statutory threshold level. In this paper, we compare three error functions for training multi-layer perceptron neural networks models of atmospheric pollution that attempt to address these requirements.

2 Neural Models of Air Pollution Time-Series

The parameters of a neural network model, $\boldsymbol{w}$, are normally determined by some form of gradient descent optimisation of an appropriate error function, $E_{\mathcal{D}}$, over

a set of labelled training examples,

$$\mathcal{D} = \{(\boldsymbol{x}_i, t_i)\}_{i=1}^{\ell}, \quad \boldsymbol{x}_i \in \mathcal{X} \subset \mathbb{R}^d, \quad t_i \in \mathcal{T} \subset \mathbb{R},$$

where $t_i \in (0,\ 1)$ for prediction of exceedences of statutory thresholds and $t_i \in \mathbb{R}^+$ for prediction of pollutant concentrations based on a vector of meteorological and other input variables $\boldsymbol{x}_i$. It has often been observed that simple maximum-likelihood estimates for the parameters of complex models often lead to severe over-fitting of the training data. In order to overcome this problem, we use instead a regularised error function, adding a term $E_{\mathcal{W}}$ penalising overly-complex models,

$$M = \alpha E_{\mathcal{W}} + \beta E_{\mathcal{D}},$$

where α and β are regularisation constants controlling the bias-variance trade-off. In this study we adopt the Bayesian regularisation scheme due to Williams [1], using a Laplace prior, i.e.

$$E_{\mathcal{W}} = \sum_{i=1}^{N} |w_i|,$$

in which the regularisation parameters α and β are integrated out analytically in the style of Buntine and Weigend [2]. An added advantage of the Laplace prior, rather than the usual Gaussian weight decay, is that redundant weights are set exactly to zero and can be pruned from the network. In the remainder of this section, we consider three data misfit terms, $E_{\mathcal{D}}$, for use in predicting episodes of poor air quality.

2.1 Logistic Regression

Logistic regression provides perhaps the most straight forward approach to predicting exceedences of statutory threshold concentrations. Assuming the target patterns, t_i, are an independent and identically distributed (i.i.d) sample drawn from a Bernoulli distribution ($t_i = 1$ indicates an exceedence, $t_i = 0$ indicates no exceedence), conditioned on the corresponding input vectors, $\boldsymbol{x}_i$, minimisation of the familiar cross-entropy error metric given by

$$E_{\mathcal{D}} = -\sum_{i=1}^{\ell} \{t_i \log y_i + (1 - t_i) \log(1 - y_i)\} \tag{1}$$

corresponds to maximisation of the likelihood of the data $\mathcal{D}$. The output layer activation function is taken to be the logistic function, $g(a) = 1/(1 + \exp\{-a\})$, restricting the output of the model to lie in the range $(0,\ 1)$. Under these conditions the output of the model is a penalised maximum likelihood estimate of the Bayesian *a-posteriori* probability of an exceedence. Unfortunately, this error metric cannot be used to obtain a direct forecast of the concentration of a given pollutant, but only an indication of the likelihood this this concentration exceeds a fixed threshold.

2.2 Conventional Sum-Of-Squares Regression

The sum-of-squares metric, $E_{\mathcal{D}} = \sum_{i=1}^{\ell}(t_i - y_i)^2$, with a linear output layer activation function corresponds to penalised maximum likelihood estimation of the conditional mean of the target values, assuming a Gaussian noise process with constant variance. This model can be simply extended to give the probability of an exceedence: The maximum likelihood estimate for the variance of the (Gaussian) target distribution is given by

$$\sigma^2 = \frac{1}{\ell}\sum_{i=1}^{\ell}(t_i - y_i)^2.$$

The probability that the observed concentration, c, exceeds a given threshold level, C, is then given by integrating the upper tail of the Gaussian probability density function, i.e.

$$p(c > C \mid \boldsymbol{x}) = \int_C^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{\frac{[z - y(\boldsymbol{x})]^2}{2\sigma^2}\right\} dz.$$

2.3 Heteroscedastic Sum-of-Squares Regression

A heteroscedastic regression model relaxes the assumption that the variance of the noise process is constant, and so attempts to estimate both the conditional mean and variance of the target distribution (e.g. Nix and Weigend [3], Williams [4]). For a Gaussian noise process, the network then has two output units, y^{μ}, estimating the conditional mean of the target distribution and y^{σ} estimating the conditional standard deviation. The negative logarithm of the likelihood is then given by

$$E_{\mathcal{D}} = -p(\mathcal{D} \mid \boldsymbol{w}) = \frac{1}{2}\sum_{i=1}^{N}\left\{\frac{[t_i - y_i^{\mu}]^2}{(y_i^{\sigma})^2} + \log(y_i^{\sigma})^2 + \log 2\pi\right\},$$

which can be used to form a penalised maximum likelihood error metric as before. Again, the probability of an exceedence is given by the integral of the probability density function of the noise process above the exceedence threshold, replacing the constant variance σ by an input-dependent variance $\sigma(\boldsymbol{x})$.

3 Results

The error metrics given in the previous section are applied to the task of prediction of episodes of poor air quality in urban Belfast due to SO_2. The target data provide hourly measurements of SO_2 taken from a Belfast monitoring station over the years 1993-1996. For any prediction of SO_2 concentration to be of practical use it must be made at least a day in advance, hence the explanatory variables include an autoregressive component beginning no later than $(T - 24)$

hours, where T denotes the time in hours at which a prediction is required. Other input variables include meteorological variables such as temperature, wind speed/direction and visibility, and day of the week, Julian day, and hour of the day. Since a Gaussian distribution is inconsistent with the observed data (being strictly positive), the final target values are the logarithm of the observed SO_2 concentrations. Hence the mean prediction models of non-linear heteroscedastic Gaussian (NLG), non-linear sum-of-squares (NLS) and linear sum-of-squares (LS) are fitting maximum likelihood estimates for a log-normal distribution, while the the non-linear logistic (NLG) and linear logistic (LL) are fitting maximum likelihood estimates for a Bernoulli distribution. An exceedence is said to have occurred if the hourly mean SO_2 concentration is greater than $350\mu gm^{-3}$. In each case, the performance statistics are computed using a four-fold cross-validation procedure, where the disjoint test partitions used in each trial are defined by the year in which the observations were made.

3.1 ROC Analysis

The *Receiver operating characteristic* (ROC) of a classifier graphically displays the trade-off between false negative (1− true positive) and false positive rates obtained by varying some parameter of the model. In this case the parameter varied is the threshold probability above which an exceedence is predicted. The area under the ROC curve gives an indication of the effectiveness of a classifier, assuming that nothing is known about the optimal ratio of misclassification costs, unity being optimal. Table 1 gives the area under the ROC curve and rankings for each model. Note the linear sum-of-squares and linear logistic regression models are both poorly calibrated in that both consistently underestimate the probability of an exceedence, however this shortcoming is not revealed by the ROC diagram.

Table 1. Area under the ROC curve for models considered.

Model	Area under ROC	Rank
NLG	0.9405	5
NLS	0.9511	4
NLL	0.9605	2
LS	0.9576	3
LL	0.9625	1

3.2 Log-likelihood Analysis

Table 2 shows the log-likelihood computed over cross-validation test partitions for the models considered. The likelihood for the task of predicting the occurrence of an exceedence, given in the second column, are calculated using the the

cross-entropy metric to evaluate the accuracy of estimates of probability of an exceedence. The likelihoods for the prediction of the concentration of SO_2 are computed using the error metrics given in the previous section.

Table 2. Log-likelihoods for considered models.

Model	log-likelihood (occurrence)	Rank	log-likelihood (prediction)	Rank
NLG	-1016.2 (-951.9)	5 (4)	-213607 (-33274)	3 (1)
NLS	-992.0	4 (5)	-34597	1 (2)
NLL	-809.1	1 (1)	*	*
LS	-916.0	3 (3)	-36316	2 (3)
LL	-841.7	2 (2)	*	*

As expected, the NLL and LL models out perform the other models in prediction of exceedences, being free of distributional assumptions regarding the noise process contaminating observations of SO_2 concentrations. Another interesting feature is the relatively poor performance of the NLG model compared to the less flexible NLS and LS models, however if the heteroscedastic variance structure is ignored and the predicted mean values used along with usual sum-of-squares estimate for the homoscedastic variance, the NLG model provides a significantly improved log-likelihood (shown in parentheses in Table 2). This is likely to be due to the observation that maximum-likelihood estimates of variance are biased since over-fitting in the model of the conditional mean reduces the apparent variance of the noise process.

3.3 McNemar's Test

Given two classifiers A and B, which classify each data point either correctly or incorrectly, McNemar's test [5] decides whether the the number of occasions that A is correct and B is incorrect is essentially the same as the number of occasions on which A is incorrect and B is correct. Table 3 gives the probabilities of the paired classifiers being essentially the same for each of the possible pairings. The lower triangle of the table gives the better classifier by this system for each pair. A (conservative) Bonferroni adjusted significance level of 0.005 is used to ensure a final significance level of 0.05 over all tests, and so there is no evidence that any of the models predict exceedences more accurately than any other.

4 Summary

In this paper, non-linear logistic regression models have demonstrated the best performance for the task of predicting episodes of poor air quality in Belfast due

Table 3. McNemar's test for considered models.

	non-linear Gaussian (NLG)	**non-linear sse (NLS)**	**non-linear logistic (NLL)**	**linear sse (LS)**	**linear logistic (LL)**
NLG	1	0.0535	0.0614	0.8750	0.2190
NLS	NLG	1	1.0000	0.1010	0.6450
NLL	NLG	NLS	1	0.0966	0.4890
LS	NLG	LS	LS	1	0.2120
LL	NLG	LL	LL	LS	1

to $S0_2$, although the differences in performance between classifiers are not statistically significant. None of the methods investigated however, provide a reliable predictor for exceedences of the statutory threshold concentration, assuming that the costs of false-positive and false-negative errors are equal. The non-linear heteroscedastic regression model provides the best estimate of the conditional mean of the concentration of SO_2. A further advantage of these models is that they can be doubly calibrated; not only is it possible to determine the probability that a pollutant exceeds a fixed threshold, accommodating changes in the costs of false-positive and false-negative errors, but also these models can still be used following a change in threshold level, due perhaps to the introduction of more stringent legislation.

5 Acknowledgements

This work was supported by the European Commission, grant number IST-99-11764, as part of its Framework V IST programme.

References

[1] Peter M. Williams. Bayesian regularisation and pruning using a Laplace prior. *Neural Computation*, 7(1):117–143, 1995.

[2] Wray L. Buntine and Andreas S. Weigend. Bayesian back-propagation. *Complex Systems*, 5:603–643, 1991.

[3] D. A. Nix and A. S. Weigand. Learning local error bars for nonlinear regression. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing*, volume 7, pages 489–496. MIT Press, 1995.

[4] Peter M. Williams. Using neural networks to model conditional multivariate densities. *Neural Computation*, 8:843–854, 1996.

[5] I. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12:153–157, 1947.

[6] D. P. Mandic and J. A. Chambers. *Recurrent neural networks for prediction - learning algorithms, architectures and stability.* Wiley series on adaptive and learning systems for signal processing, communications and control. John Wiley & Sons Ltd., Chichester, 2001.