Withdrawing an example from the training set: an analytic estimation of its effect on a non-linear parameterised model

Gaétan MONARI*′**, Gérard DREYFUS*

*École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris

Laboratoire d'Électronique

10, rue Vauquelin - F 75005 PARIS - FRANCE

**USINOR

DSI/DISA SOLLAC FOS bat. LB1

F 13776 FOS-sur-Mer Cedex - FRANCE

Abstract

For a non-linear parameterised model, the effects of withdrawing an example from the training set can be predicted. We focus on the prediction of the error on the left-out example, and of the confidence interval for the prediction of this example. We derive a rigorous expression of the first-order expansion, in parameter space, of the gradient of a quadratic cost function, and specify its validity conditions. As a consequence, we derive approximate expressions of the prediction error on a given example, and of the confidence interval thereof, had this example been withdrawn from the training set. We show that the influence of an example on the model can be summarised by a single parameter. These results are applicable to leave-one-out cross-validation, with a considerable decrease in computation time with respect to conventional leave-one-out. The paper focuses on the theoretical aspects of the question; both academic illustrations and large-scale industrial examples are described in [9].

1 Introduction

When performing non-linear regression with a family of parameterised functions that are non-linear with respect to their parameters, such as neural networks, an approximate expression of the least-squares solution may be derived by using a Taylor expansion, *in parameter space*, of the model in the vicinity of the minimum of the cost function. Unlike previous derivations of this expansion, the derivation presented here is exact, so that the assumptions under which this expansion is valid can be clearly stated. Then, we show that this result can be used for

predicting the effect of withdrawing an example from the training set on its prediction and on the corresponding confidence interval.

The present paper discusses static single-output processes with a non-random input vector $\boldsymbol{x}$ and an output y_p, which is considered as a measurement of a random variable Y_p. We assume that an appropriate model can be written under the form $y_p = r(\boldsymbol{x}) + w$, where w is a zero-mean random variable and $r(\boldsymbol{x})$ is the unknown regression function. A family of parameterised functions $f(\boldsymbol{x}, \theta)$ is used to approximate r and a data set of N input-output pairs $\{\boldsymbol{x}^k, y_p^k\}_{k=1,...,N}$ is assumed to be available for estimating the parameters of the model. In the following, all vectors are column vectors, denoted by boldface letters, e.g. the vectors $\boldsymbol{x}$ and $\{\boldsymbol{x}^k\}$.

2 Local approximation of the least-squares solution

We denote by θ_{LS} the least-squares solution: a q-vector of parameters that minimises the quadratic cost function $J(\theta) = {}^t[\boldsymbol{y_p} - \boldsymbol{f}(X, \theta)]\,[\boldsymbol{y_p} - \boldsymbol{f}(X, \theta)]$, i.e. that cancels its derivative[1].

If the model is not linear with respect to its parameters, an approximate solution can be obtained by using a Taylor expansion of the model in the vicinity of a parameter vector θ^* in parameter space. In order to obtain a *valid* first-order development of $\frac{\partial J}{\partial \theta}$ in $(\theta - \theta^*)$, one has to resort to a *second-order* development of $\boldsymbol{f}(X, \theta)$ in the vicinity of θ^*:

$$\boldsymbol{f}(X, \theta) \cong \boldsymbol{f}(X, \theta^*) + Z\,(\theta - \theta^*) + {}^t(\theta - \theta^*)\,\boldsymbol{S}\,(\theta - \theta^*) \qquad (1)$$

In the previous relation:

- $Z = {}^t[\boldsymbol{z}^1, ..., \boldsymbol{z}^N]$ stands for the jacobian matrix of the model, where $\boldsymbol{z}^i = \left.\frac{\partial f(\boldsymbol{x}^i, \theta)}{\partial \theta}\right|_{\theta = \theta^*}$.
- $\boldsymbol{S} = \sum_{i=1}^{N} S(\boldsymbol{x}^i)\,\boldsymbol{e}_i$ is a third-order tensor in which $S(\boldsymbol{x}^i)$ is a (q, q) matrix defined as $S(\boldsymbol{x}^i) = \left(\left.\frac{\partial^2 f(\boldsymbol{x}^i, \theta)}{\partial \theta_j\, \partial \theta_k}\right|_{\theta = \theta^*}\right)_{\substack{j=1,...,q \\ k=1,...,q}}$ and $\boldsymbol{e}^i$ is the i^{th} vector of the orthonormal basis of $\Re^N$.

Using (1) in the cost function, one obtains, after differentiation, and neglecting the terms whose order in $(\theta - \theta^*)$ exceeds 1:

[1] We denote $\boldsymbol{f}(X, \theta) = {}^t[f(\boldsymbol{x}^1, \theta), ..., f(\boldsymbol{x}^N, \theta)]$, where X is the (N, n) matrix ${}^t[\boldsymbol{x}^1, ..., \boldsymbol{x}^N]$.

$$\frac{\partial J}{\partial \theta} = \frac{\partial J}{\partial (\theta - \theta^*)} \cong -2\,{}^tZ\left(\boldsymbol{y}_p - \boldsymbol{f}(x, \theta^*)\right) + \left\{2\,{}^tZZ - 2\sum_{i=1}^{N}\left(y_p^i - f(\boldsymbol{x}^i, \theta^*)\right) S(\boldsymbol{x}^i)\right\}(\theta - \theta^*) \qquad (2)$$

The (q, q) matrix between the curled brackets is the Hessian of the cost function, that is $H = \left(\frac{\partial^2 J(\theta)}{\partial \theta_j \, \partial \theta_k}\right)_{\substack{j = 1, ..., q \\ k = 1, ..., q}}$. If the second term of the Hessian can be neglected, one can approximate H by $2\,{}^tZ\,Z$ (Levenberg-Marquardt approximation). In all the following, we assume that Z has full rank[2].

Finally, after cancelling the gradient of J, one obtains an approximate expression of θ_{LS}:

$$\theta_{LS} \cong \theta^* + ({}^tZ\,Z)^{-1}\,{}^tZ\,[\boldsymbol{y}_p - \boldsymbol{f}(X, \theta^*)] \qquad (3)$$

Relation (3) is valid under two assumptions:

- the second-order terms of the Taylor expansion of $\frac{\partial J}{\partial \theta}$ are negligible with respect to first-order terms, i.e. the cost function can locally be approximated by a quadratic surface; estimations of the curvature of the cost function have been investigated by several authors (see for instance [11], [1]),
- the Hessian of the cost function can be approximated by $2\,{}^tZ\,Z$; the latter approximation is discussed, in a different context, in [2] and [7].

In the case of a model that is linear with respect to the parameters, relation (3) is an equality.

This result was obtained previously by other authors, starting from a *first-order* development of $\boldsymbol{f}$: in so doing, they actually overlooked the second term of H instead of neglecting it knowingly (see for instance [11], where the result is used for estimating confidence intervals for the model parameters and for the model output).

3 The effect of withdrawing an example from the training set

It this section, the above approximation of the least-squares solution is used to predict the effects of withdrawing an example from the training set.

In all the following, the quantities pertaining to the models that have been trained on all examples except example i are denoted with a superscript $^{(-i)}$: $\boldsymbol{f}^{(-i)}(X, \theta)$ and $\boldsymbol{y_p}^{(-i)}$ are $(N - 1)$

[2] Indeed, a rank deficiency of Z indicates an under-determination of some coefficients, which means that the corresponding model is overfitted (cf. [9]): such models should be discarded.

vectors, and $Z^{(-i)}$ is a (N - 1, q) matrix. Conversely, quantities that have no superscript pertain to models trained with all examples.

3.1 The effect of withdrawing an example on its prediction

Assuming that the withdrawal of example i from the training set generates but a small variation of the least-squares solution, one can obtain an approximate expression of $\theta_{LS}^{(-i)}$ in the vicinity of θ^*, as for relation (3):

$$\theta_{LS}^{(-i)} \cong \theta^* + \left({}^t Z^{(-i)}\, Z^{(-i)}\right)^{-1} {}^t Z^{(-i)}\, [\boldsymbol{y}_p^{(-i)} - \boldsymbol{f}^{(-i)}(X,\, \theta^*)] \tag{4}$$

Combining (3) and (4), one obtains, by using an appropriate matrix inversion lemma (see for instance [1]):

$$\theta_{LS}^{(-i)} \cong \theta_{LS} - \left({}^t Z\, Z\right)^{-1} \boldsymbol{z}^i \frac{R_i}{1 - h_{ii}} \tag{5}$$

where $\boldsymbol{z}^i$ is the vector whose components are the i^{th} column of matrix Z, R_i is the residual for example i: $R_i = y_{pi} - f(\boldsymbol{x}^i,\, \theta_{LS}) = y_{pi} - f(\boldsymbol{x}^i,\, \theta^*) - {}^t\boldsymbol{z}^i\, \theta_{LS}$ and

$$h_{ii} = {}^t\boldsymbol{z}^i\, ({}^t Z\, Z)^{-1}\, \boldsymbol{z}^i \tag{6}$$

is the *leverage* of example i, i.e. the i^{th} component of the projection, on the solution subspace, of the unit vector along axis i. By definition, all $\{h_{ii}\}_{i=1,\ldots,N}$ lie between 0 and 1, and satisfy the following relation:

$$\sum_{i=1}^{N} h_{ii} = q \tag{7}$$

We denote by $R_i^{(-i)}$ the error on the prediction of example i when it is withdrawn from the training set: $R_i^{(-i)} = y_{p\,i} - f(\boldsymbol{x}^i,\, \theta^*) - {}^t\boldsymbol{z}^i\, \theta_{LS}^{(-i)}$. The expression of $R_i^{(-i)}$ as a function of R_i is obtained immediately: $R_i^{(-i)} \cong R_i + {}^t\boldsymbol{z}^i \left(\theta_{LS} - \theta_{LS}^{(-i)}\right)$. Using relation (5), the same expression as in the linear case (see [5]) is readily derived:

$$R_i^{(-i)} \cong \frac{R_i}{1 - h_{ii}} \tag{8}$$

Similarly, an approximation of the quadratic cost function is obtained:

$$J(\theta_{LS}^{(-i)}) \cong J(\theta_{LS}) - \frac{R_i^2}{1 - h_{ii}} \tag{9}$$

A similar idea was proposed in [6] and [12]. However, the derivation of their relations is not correct, as evidenced by the fact that their results are not exact in the linear case, as opposed to relations (5), (8) and (9).

An illustration of the accuracy of the previous approximations, on problems of various sizes, can be found in [9].

3.2 The effect of withdrawing an observation on the confidence interval of its prediction

In [11], an approximate expression for a confidence interval on the output Y_p of a non-linear model is derived, under suitable assumptions, with a confidence level 1 - α.

For observation *i* of the training set, it can be written as:

$$E\left(Y_p \left| \boldsymbol{x}^i \right.\right) \in f\left(\boldsymbol{x}^i, \theta_{LS}\right) \pm t_\alpha^{N-q}\, s \sqrt{{}^t\boldsymbol{z}^i \left({}^tZZ\right)^{-1} \boldsymbol{z}^i} = f\left(\boldsymbol{x}^i, \theta_{LS}\right) \pm t_\alpha^{N-q}\, s \sqrt{h_{ii}}\,, \qquad (10)$$

where t_α^{N-q} is the value of a Student variable with N - q degrees of freedom and a level of significance 1 - α and s is an estimation of the residual standard deviation of the model.

Relation (10) shows that the analytic approach of confidence intervals involves the same quantity (h_{ii}) as the prediction of the effect of withdrawing an example from the training set; this is not surprising since both groups of relations stem from the Taylor expansion of the model in parameter space. This is why we have chosen to focus on these confidence intervals, among the many confidence intervals that have been proposed for non-linear models ([13]).
As in the previous section, it is possible to estimate the confidence interval associated with the prediction of a withdrawn example: given an input vector $\boldsymbol{x}^i$, the approximate confidence interval, with confidence level 1 - α, for the model obtained after withdrawing example *i* from the training set, is given by:

$$E^{(-i)}\left(Y_p \left| \boldsymbol{x}^i \right.\right) \in f\left(\boldsymbol{x}^i, \theta_{LS}^{(-i)}\right) \pm t_\alpha^{N-q-1}\, s^{(-i)} \sqrt{{}^t\boldsymbol{z}^i \left({}^tZ^{(-i)} Z^{(-i)}\right)^{-1} \boldsymbol{z}^i} \qquad (11)$$

Using the same matrix inversion lemma as for relation (5), and combining relations (8) and (11), it can easily be shown that:

$$E^{(-i)}\left(Y_p \left| \boldsymbol{x}^i \right.\right) \in f\left(\boldsymbol{x}^i, \theta_{LS}\right) - \frac{h_{ii}}{1-h_{ii}} R_i \pm t_\alpha^{N-q-1}\, s^{(-i)} \sqrt{\frac{h_{ii}}{1-h_{ii}}}\,. \qquad (12)$$

In general, $s^{(-i)}$ can be approximated by s.

3.3 Interpretation of the leverages

All previous relations show that the leverages $\{h_{ii}\}_{i=1,\ldots,N}$ play a crucial role.

Two limit cases are of interest:

- if axis i is orthogonal to the solution subspace, defined by the columns of matrix Z, all columns have their i^{th} component equal to zero; hence $\boldsymbol{z}^i = 0$ and $h_{ii} = 0$. Example i has essentially no influence on the model (it has none in the case of a linear model), as evidenced by relations (5), (8) and (9); such a situation cannot arise if the model has a constant input;
- if axis i lies within the solution subspace, then $h_{ii} = 1$ and $R_i \cong 0$. In other terms, example i is learnt almost perfectly (it is learnt perfectly in the case of a linear model).

Thus, the leverage h_{ii} is a way of estimating the influence of example i on the model. The closer h_{ii} to 1, the larger the influence of example i*:* as expected, the confidence interval on its prediction, if it is withdrawn from the training set, becomes infinite if h_{ii} approaches 1 (relation (12)). This means that the model has made use specifically of some of its parameters in order to fit this example almost perfectly (R_i becomes very small).

Note that relation (7) can be interpreted as follows: since the $\{h_{ii}\}_{i=1,\ldots,N}$ sum to q, i.e. to the number of degrees of freedom of the model, $\frac{h_{ii}}{q}$ can be regarded as the fraction of the available parameters that are used to fit example i: a leverage close to 1 means that almost one degree of freedom has been used to fit example i; conversely, a leverage close to zero means that almost no parameter has been made use of in order to fit this example, i.e. that example i has no influence on the model. If all $\{h_{ii}\}_{i=1,\ldots,N}$ are equal to $\frac{q}{N}$, a fraction $\frac{1}{N}$ of the degrees of freedom of the model is devoted to fitting each example.

In the present paper, we have made use of the diagonal elements h_{ii} of the leverage matrix $H = {}^tZ\,({}^tZ\,Z)^{-1}\,Z$. The off-diagonal terms of this matrix can be used for estimating the effect of the deletion of pairs of examples from the training set, as investigated, in the linear case, in [8], using Cook's distance [4]. The extension to the nonlinear case, along the same lines as discussed above, is still an open problem.

4. Conclusion

We have given a rigorous derivation of the first-order expansion, in parameter space, of the gradient of the cost function for a model that is not linear with respect to its parameters, and we have stressed the assumptions under which these approximations are valid. From this result, we have derived the effect of withdrawing an example from the training set on the prediction of the model and on its confidence interval.

These results can be applied to leave-one-out or jackknife procedures, thereby exempting the designer altogether from training as many models as examples. In [9] and [10], the application of these techniques to model selection, and their limitations, are investigated in detail. Model selection is performed in two steps: (i) for each candidate architecture, selection of a minimum of the cost function (after training on the whole set of available data), on the basis of a *predicted* leave-one-out score, obtained by replacing, in the traditional leave-one-out score, the actual error on the left out example by the predicted error given by relation (8); (ii) among the resulting models, select the model for which the distribution of the leverages is most peaked around $\frac{q}{N}$. Thus, starting from the leave-one-out principle and making use of the confidence intervals, the proposed method is a new, simple and computationally inexpensive (as opposed to Bayesian methods), way of dealing with overfitting by monitoring the influence of each example on the model; in addition, the stability problems of the leave-one-out technique (see [3]) are overcome. The technique has been successfully applied to an industrial problem: the on-line prediction of the quality of spot welds from measurements performed during welding.

Acknowledgements

This research was performed with the support of SOLLAC (Group USINOR) and of ANRT.

The authors wish to thank Prof. S. Canu for drawing their attention to the validity assumptions of the Taylor expansion. They are grateful to Dr. L. Personnaz and Dr. I. Rivals for helpful discussions, and for providing a derivation of relation (8) from relation (5).

References

[1] A. Antoniadis, J. Berruyer, R. Carmona, Régression non linéaire et applications, Economica, Paris (1992).

[2] C.M. Bishop, Neural Networks for Pattern Recognition, Third Edition, Clarendon Press, Oxford (1997).

[3] L. Breiman, Heuristics of Instability and Stabilization in Model Selection, Annals of Statistics 24 (1996) 2350-2383.

[4] R.D. Cook and S. Weisberg, Characterization of an empirical influence function for detecting influential cases in regression, Technometrics 22 (1980) 495-508.

[5] B. Efron and R.J. Tibshirani, An Introduction to the Boostrap, Chapman & Hall, New-York (1993).

[6] L.K. Hansen and J. Larsen, Linear Unlearning for Cross-Validation, Advances in Computational Mathematics 5, (1996) 269-280.

[7] B. Hassibi and D.G. Stork Second-order derivatives for network pruning: optimal brain surgeon, S.J. Hanson, D.J. Cowan & C.L. Giles (Eds), Advances in Neural Information Processing Systems 5 (1993) 164-171.

[8] A.J. Lawrance, Deletion, influence and masking in regression, J. R. Statist. Soc. B 57 (1995) 181-189.

[9] G. Monari (1999), Sélection de modèles non linéaires par leave-one-out: étude théorique et application des réseaux de neurones au procédé de soudage par points, Thèse de l'Université Paris 6, available from:

http://www.neurones.espci.fr/Francais.Docs/publications.html

[10] G. Monari & G. Dreyfus, "Overfitting Monitoring by Leverage Limitation", in preparation.

[11] G.A.F Seber and C.J. Wild, Nonlinear regression, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York (1989).

[12] P.H. Sorensen, M. Norgard, L.K. Hansen and J. Larsen, Cross-Validation with LULOO, Proceedings of the International Conference on Neural Information Processing - ICONIP '96 (1996).

[13] R.J. Tibshirani, A Comparison of Some Error Estimates for Neural Models, Neural Computation 8 (1996) 152-163.