Fast Bootstrap for Model Structure Selection

A. Lendasse, V. Wertz
Cesame, UCL, av. Georges Lemaître 4
1348 Louvain-la-Neuve, Belgium
email: {lendasse, wertz}@auto.ucl.ac.be

G. Simon, M. Verleysen
DICE, UCL, Place du Levant 3,
1348 Louvain-la-Neuve, Belgium
email : {gsimon, verleysen}@dice.ucl.ac.be

1 Introduction

In this paper we propose an effective procedure to reduce the computation time of a bootstrap approximation of the generalization error in a family of nonlinear regression models. The bootstrap [1] is based on the general plug-in principle which permits to obtain an estimator of a statistic according to an empirical distribution. In our context we use the bootstrap to estimate the generalization error of several models in order to choose the "best" one. The bootstrap estimator is computed over a finite number N of new samples x^* generated from the original sample x by drawing with replacement. The bootstrap estimate of the generalization error is given by

$$\hat{e}_{gen} = e_{app} + optimism, \tag{1}$$

where e_{app} is the apparent error obtained when evaluating the model built (learned) on the original sample x on the same sample (learning error), and optimism is a correction term aiming to estimate the difference between a learning and a generalization error. The optimism is computed on the N bootstrap replications:

$$optimism = E[e_{x^*}(\hat{F}_x) - e_{x^*}(\hat{F}_{x^*})], \tag{2}$$

where $E[]$ is the statistical expectation computed over all bootstrap replications and is the error for a model developed (learned) on the x^* sample and evaluated on the empirical distribution. and are the empirical distribution functions in the real world and in the bootstrap world respectively.

2 Methodology

It is well known that the apparent error e_{app} of a nonlinear regression model is usually roughly exponentially or quadratically decreasing with the number p of parameters in the model. With a good approximation, e_{app} can thus be expressed as one of the following expressions:

$$e_{app} \approx Ae^{-Bp} \quad or \quad e_{app} \approx \frac{1}{Ax + Bx^2} \tag{3}$$

A second empirical fact is that the *optimism* increases roughly linearly with the number p of parameters, leading to:

$$optimism \approx Cp + D. \tag{4}$$

The bootstrap estimate of the generalization error would then give, for some parameters A, B, C and D:

$$\hat{e}_{gen} = Ae^{-Bp} + Cp + D \; or \; = \frac{1}{Ax + Bx^2} + Cp + D$$

The principle of the method is then to make a limited number of experiments to estimate A, B, C and D. A and B are evaluated by (3) with models (with different values of p) using the original sample x both for learning and test. The C and D values can be computed according to (4) with models built on bootstrap replicates x^* and evaluated on both the original sample x and the bootstrap samples x^*.

3 Experimental Results

We illustrate the method described in the previous section on a standard benchmark in time-series prediction: the Santa Fe A time series. The model we used is chosen a priori to be:

$$\hat{y}(t+1) = f(y(t),y(t-1),...,y(t-6)).$$

A RBFN is characterised by its number n of Gaussian kernels (or hidden units). We trained 7 RBFNs on the Santa Fe learning dataset, for n = 20, 40, 60, 80, 100, 120 and 140 respectively. With the apparent generalization error obtained for those values of n we can compute A and B (in the least mean square sense) and deduce an estimate of the apparent error. Then bootstrap estimates of the optimism are evaluated for the same values of n. A linear interpolation gives C and D. The minimum of the generalization error is attained for n = 103 (results are presented in Fig.1). Then we repeat the same experiment with four different n instead of seven, with values of n equals 20, 60, 100 and 140 respectively. The optimal n found in this case is a coherent result of 102. The optimal n found with a huge test set is n = 100.

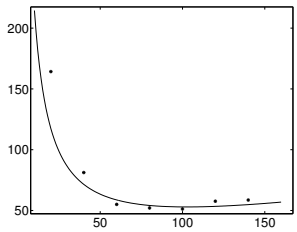

FIG. 1 – *$\hat{e}_{gen}$ with regard to n in solid line, bootstrap estimates are marked with asterisks*

References

[1] Efron, B. and Tibshirani, R. J., An introduction to the bootstrap. Chapman & Hall, New York , 1993.

Aknowledgement. G. Simon is funded by the Belgian F.R.I.A. M. Verleysen is Senior Research Associate of the Belgian F.N.R.S. The work of A. Lendasse and V. Wertz is supported by the Interuniversity Attraction Poles, initiated by the Belgian Federal State, Ministry of Sciences, Technologies and Culture. The scientific responsibility rests with the authors.