Note on free lunches and cross-validation

Cyril Goutte
Department of Mathematical Modelling, building 321
Technical University of Denmark, DK-2800 Lyngby, Denmark

January 28, 1997

Abstract. The "no free lunch" theorems (Wolpert and Macready, 1995) have sparked heated debate in the computational learning community. A recent communication, (Zhu and Rohwer, 1996) attempts to demonstrate the inefficiency of cross-validation on a simple problem. We elaborate on this result by considering a broader class of cross-validation. We show that when used more strictly, cross-validation can yield the expected results on simple examples.

1 Introduction

A recent contribution to computational learning, and neural networks in particular, the "no free lunch" (NFL) theorems (Wolpert and Macready, 1995) give surprising insight into computational learning schemes. One implication of the NFL is that no learning algorithm performs better than random guessing over all possible learning situation.

This means in particular that the widely used *cross-validation* (CV) methods, if successful in some cases, should fail on others. Considering the popularity of these schemes, it is therefore of considerable interest to exhibit such a problem where the use of CV leads to a decrease in performance. (Zhu and Rohwer, 1996) propose a simple setting in which a "cross-validation" method yields worse results than the maximum likelihood estimator it is based on. In the following, we extend these results to a stricter definition of cross-validation and provide an analysis and discussion of the results.

2 Experiments

The experimental setting is the following: a Gaussian variable x has mean μ and unit variance. The mean should be estimated from a sample of n realisation of x. Three estimators are compared:

A(n): The mean of the n point sample. It is the maximum likelihood and least mean squares estimator, and also the optimal unbiased estimator.

B(n): The maximum of the n point sample.

C(n): The estimator obtained by a cross-validation choice between A(n) and B(n).

The original "cross-validation" setting of (Zhu and Rohwer, 1996) will be noted Z($n+1$): it samples *one additional point*, and chooses the estimator A or B that is closer to this point. However, the widely used concept of cross-validation (Ripley, 1996; FAQ, 1996) corresponds to *resampling* and *averaging* estimators, rather than sampling additional points. In this context, the estimator proposed in (Zhu and Rohwer, 1996) is closer to the "split-sample" a.k.a. "hold out" method. This method is known to be noisy, especially when the validation set is small, which is the case here.

On the other hand, a thorough cross-validation scheme would use several validation sets resampled from the data and average over them before choosing estimator A or B. In the *leave-one-out* (LOO) flavour, the CV score is calculated as the average distance between each point and the estimator obtained on the rest.

Note that in this setting, estimator Z operates with more information (one supplementary data point) than the LOO estimator. The result of an experiment done with $n = 16$ and 10^6 samples gives mean squared errors:

$$A(16) : 0.0624 \qquad B(16) : 3.4141 \qquad C_{LOO}(16) : 0.0624 \qquad Z(16+1) : 0.5755 \tag{1}$$

In this case, it seems that the proper cross-validation procedure *always* picks estimator A, whose theoretical mean squared error is $1/16 = 0.0625$.

3 Short analysis

The simple setting used in these experiments allows for a full analysis of the behaviour of C_{LOO}. Consider n data points x_i. The *leave-one-out* cross-validation estimate is computed by averaging the squared error between each point and the average of the rest. Let us denote by $\overline{x}$ the average of all x_i and by $\overline{x}_k$ the average of all x_i *excluding* example x_k: $\overline{x}_k = \frac{n\overline{x}-x_k}{n-1}$. Accordingly, $(\overline{x}_k - x_k) = \frac{n}{n-1}(\overline{x} - x_k)$, hence the final cross-validation score for estimator A:

$$\overline{CV} = \left(\frac{n}{n-1}\right)^2 S\left(\overline{x}\right) \tag{2}$$

where $S\left(w\right) = \frac{1}{n}\sum_{i=1}^{n}\left(w - x_i\right)^2$ is the mean squared error between estimator w and the data. Let us now note $x^* = \max\left\{x_i\right\}$ the maximum of the (augmented) sample, and $x^{**} = \max\left(\{x_i\} \setminus x^*\right)$ the second largest element. The cross-validation score for estimator B is:

$$CV^* = S\left(x^*\right) + \frac{1}{n}\left(x^* - x^{**}\right)^2 \tag{3}$$

In order to realise how the cross-validation estimator behaves, let us first recall that $\overline{x}$ is the least mean squares estimator. As such, it minimises $S\left(w\right)$. Furthermore, from Huygens' formula, $S\left(x^*\right) = S\left(\overline{x}\right) + \left(\overline{x} - x^*\right)^2$. Accordingly, we can re-write $CV^* = S\left(\overline{x}\right) + \left(\overline{x} - x^*\right)^2 + \frac{1}{n+1}\left(x^* - x^{**}\right)^2$.

These observation show that in order for (3) to be lower than (2) requires an extremely unlikely situation. x^* should be quite close to both $\overline{x}$ and x^{**}. One such situation *could* arise in the presence of a negative outlier. As the mean is not robust, the outlier would produce a severe downward bias in estimator A. Thus the maximum could very well be a better estimator than the mean in such a case. However, no such happenstance was observed in 5.10^6 experiments.

4 Discussion

1. In experiment (1), the LOO estimator does not use the additional data point alloted to C. When using this data point, the performance is identical: CV does not extract any information from the extra sample, but at least manages to keep the information in the original sample.

2. The cross-validation estimator does not perform better than A(16), and yields worse performance than A(17), for which the theoretical MSE is $1/17 \simeq 0.0588$. However, it should be pointed out that the setting imposes a *choice* between A and B. The *optimal choice* over 10^5 samples leads to a mean squared error of just 0.0617. This is a lower bound for estimator C, and is significantly beyond the optimal, 17-points estimator. On the other hand, a *random choice* between A and B leads to a mean squared error of 1.7364 on the same sample.

3. It could be objected that LOO is just one more flavour of cross-validation, so results featuring this particular estimator do not necessarily have any relevance to CV methods as a whole. Let us then compare the performance of m-fold CV on the original 16 point sample. It consists in dividing the sample into m validation sets, and averaging the validation performance of the m estimators obtained on the remaining data. For 10^5 samples, we get (CV_m is the m-fold CV estimator):

Estimator	A	B	$CV_{16} = C_{LOO}$	CV_8	CV_4	CV_2
MSE	0.0624	3.4141	0.0624	0.0624	0.0624	0.0628

The slight decrease in performance in CV_2 is due to the fact that we average over only 2 validation sets. If we resample two additional sets, the performance is identical to all other CV estimators.

4. While requiring additional computation, none of the CV estimators above gain anything on A (even with the help of one additional point). Better performance can however be observed with a different choice of estimators. Consider e.g. D(n) a median of the sample, and E(n) the average between the min and the max. Using 10^5 sample, we compare the CV estimator to D and E calculated on 16 point samples:

$$D(16) : 0.0949 \qquad E(16) : 0.1529 \qquad C_{LOO}(16) : 0.0931 \tag{4}$$

Here cross-validation outperforms both estimators it is based on.

5 Conclusion

The no free lunch results imply that for every situation where a learning method performs better than random guessing, another situation exists where it performs correspondingly worse. Numerous reports of successful applications of usual learning procedures suggest that the "unspecified prior" under which they outperform random guessing verify in a number of practical cases. Exhibiting these assumptions is of importance in order to check whether the conditions for success hold when tackling a new problem. However, it is unlikely that such a simple setting could challenge these yet unknown assumptions. Cross-validation has many drawbacks, and it is far from being the most efficient learning method (even among non-Bayesian frameworks). In that simple case, though, it provides perfectly decent results.

We now know that there is "no free lunch" for cross-validation. However, the task of exhibiting an easily understandable, non-degenerate case where it fails has yet to be completed. Furthermore, the task of exhibitting the "hidden" prior under which cross-validation is beneficial provides challenging prospects for the future.

Acknowledgments This work was partially supported by a research fellowship from the Technical University of Denmark (DTU). I am grateful to the (anonymous) referees for constructive criticism, and to Huaiyu Zhu and Lars Kai Hansen for challenging remarks on earlier drafts of the paper.

References

FAQ (1996). Neural network Frequently Asked Questions. URL: `ftp://ftp.sas.com/pub/neural/FAQ3.html`. FAQ in comp.ai.neural-nets, Part 3.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge University Press.

Wolpert, D. H. and Macready, W. G. (1995). The mathematics of search. Technical Report SFI-TR-95-02-010, Santa Fe Institute.

Zhu, H. and Rohwer, R. (1996). No free lunch for cross validation. *Neural Computation*, 8(7):1421–1426.