To appear in Neural Computation

No Free Lunch for Cross Validation

Huaiyu Zhu and Richard Rohwer
Neural Computing Research Group
Dept of Computer Science and Applied Mathematics
Aston University, Birmingham B4 7ET, UK
Email: zhuh@aston.ac.uk, Fax: +44 121 333 6215

February 19, 1996

Abstract

It is known theoretically that an algorithm cannot be good for an arbitrary prior. We show that in practical terms this also applies to the technique of "cross validation", which has been widely regarded as defying this general rule. Numerical examples are analysed in detail. Their implications to researches on learning algorithms are discussed.

1 Introduction

Recently, practical implications of the so-called "No Free Lunch Theorems" (NFL) by (Wolpert and Macready, 1995) have become a contentious issue among neural network researchers. For our purpose, the NFL Theorems can be summarised as the following.

- With a uniform prior, any algorithm performs as well as random guessing.
- With a "uniform hyperprior" of priors, any algorithm performs as well as random guessing.

The main implication is that if an algorithm performs better than random guessing on some prior, then it necessarily performs worse than random guessing on some other prior, by the same amount. Therefore it is meaningless to say an algorithm is good without specifying the prior.

On the other hand, it is widely believed that some frequentist statistical techniques, such as cross validation (CV), bootstrap, etc., are universally good. More specifically, it is often claimed that they will automatically discover useful structure in the problem, if there is any, and will be harmless otherwise. Although this has been theoretically shown not to be the case (Wolpert and Macready, 1995), we consider it still of great instructive value to see why and how CV fails in a numerical experiment.

2 Experiment and Short Analysis

Suppose we have a Gaussian variable x, with mean μ and unit variance. We consider the following three estimators for estimating μ from a sample of size n.

- A: The sample mean. It is optimal both in the sense of Maximum Likelihood and Least Mean Squares.
- B: The maximum of the sample. It is a bad estimator in any reasonable sense.
- C: Cross validation to choose between A and B, with one extra data point.

The numerical result with $n = 16$, averaged over 10000 samples, gives mean squared errors:

$$A : 0.0627 \qquad B : 3.4418 \qquad C : 0.5646. \tag{1}$$

This clearly shows that CV is harmful in this case, despite the fact that it is based on a larger sample.

To alleviate concern that the verdict might be caused by statistical fluctuation, we also repeated the experiment with 10^6 samples, which gives three significant digits. The result is,

$$A : 0.0625 \qquad B : 3.4137 \qquad C : 0.5754. \tag{2}$$

Note that the theoretical mean squared error for A is $1/16 = 0.0625$.

It might at first appear that this is a very artificial example, which is not what normally occurs in practice. To this we have two answers, short and long. The short answer is from first principles. Any counter-example, however artificial it is, clearly demolishes the hope that CV is a "universally beneficial method".

3 Full Analysis and Further Experiments

The longer answer is divided into several parts, which hopefully will answer any potential criticism of any aspect:

1. The CV is performed on extra data points. We are not requiring it to perform as well as the mean on 17 data points. If it cannot extract more information from the one extra data point, a minimum requirement is that it keeps the information in the original 16 points. But it cannot even do this.

2. Denote by T_a the ath percentile of the sample. Then the maximum of a sample is T_{100}. The median is T_{50}, which is in fact a quite reasonable estimator. Let us use a larger CV set (of size k), and replace estimator B with a different percentile. The result is that for CV to work, it needs $k > 2$ for the median, and $k > 16$ for T_{70} .

3. It is not true that we have set up a case in which CV cannot win because T_{100} is at the boundary of the interval spanned by the sample. There is indeed a small probability that a sample can be so bad that the sample maximum is even a better estimate than the sample mean. However to utilise such rare chances to good effect k must be at least several hundred (maybe exponential) while $n = 16$. We know that such a k exists since $k = \infty$ certainly helps. However, to adopt such a method is clearly absurd.

4. The reason CV fails in this particular case is easy to see. It is true that C will choose A for most of time, as it should be, but whenever it chooses B, it is most likely for the wrong reason. For the case at hand, in 10000 samples, about 30 are such that B is better than A. However, based on one extra point, C will prefer B for about 2000 cases. Among them, only about 15 samples are for the right reason, that B is genuinely better. For the remainder 1985 samples the reason is that the validation point happens to be closer to B. Furthermore, among the 30 worthy cases it fails to pick half of them because the validation point happens to be at the wrong side of the sample mean. As the CV size increases, the chance of picking B for the wrong reason decreases. Following is a typical run showing the role of validation set size.

validation size k	number of samples in 10000 samples		
	C prefers B	B is genuinely better	C picks B correctly
1	1989	26	13
2	1380	26	15
3	952	26	13
5	601	26	18
8	362	26	16
13	239	26	15
21	139	26	13

5. We have chosen estimator A to be the known optimal estimator in this case to make the mechanism easier to understand, but it can be replaced by something else. For example, both A and B can be some reasonable averages over percentiles, such as $A = (T_{40} + T_{60})/2$ and $B = (T_{10} + T_{50} + T_{90})/3$, so that without detailed analysis it is hard to see which is better. It may appear doing CV might generate C which is better. Such beliefs can be defeated by similar counter-examples. In most cases C will have a performance intermediate between A and B, unless the CV set is enormously larger than the original set.

6. The above may give the wrong impression that it is impossible to mix several estimators to get a better estimator, which is not true. If we have a sample of size 101, and sort it in increasing order, then each data point is a percentile, from T_0 to T_{100}. Among these 101 different estimators, the optimal is the median T_{50}. A simple arithmetic average of these estimators gives the sample mean, which is a better estimator than the best of the estimators it is based on, even without using any fresh data.

7. The above scheme of CV my appear different from what is familiar, but here is a "practical example" which shows that it is indeed what people normally do. Suppose we have a random variable which is either Gaussian or Cauchy. Consider the following three estimators (See (Fisher, 1925) for definition and properties of efficiency):

   - A: The sample mean: it has 100% efficiency for Gaussian, and 0% efficiency for Cauchy.
   - B: The sample median: it has $2/\pi = 63.66\%$ efficiency for Gaussian and $8/\pi^2 = 81.06\%$ efficiency for Cauchy.
   - C: Cross validation on an additional sample of size k, to choose between A and B.

   Intuitively it may appear reasonable to expect CV to pick out the correct one, for most of the time, so that averaging over all samples C ought to be superior to both A and B. But this is not so, since this will depend on the *prior* mixing probability of these two sub-models. If the variable is in fact always Gaussian, then we have just seen that if $n = 16$, CV will be worse than sticking with A unless $k > 2$. The same is even more true in the reversed order, since the mean is an essentially useless estimator for the Cauchy distribution. This also shows that averaging among estimators is not always a good thing to do.

8. In many application problems CV is performed on a continuous hyperparameter instead of a choice between discrete alternatives. For an example of this type consider the t distribution, which connects the Cauchy and Gaussian distributions by varying the "hyperparameter", the degrees of freedom m. Suppose we have obtained quite good estimators for each integer m, how can we obtain a good estimator if we do not know m? Cross validation can be either good or bad, depending on the prior mixing distributions among all the t-distributions. However, if we had known that, we would be better-off using Bayesian methods, which might turn out to be CV, but in this case it is highly likely to be some other estimator far better than CV.

9. These examples also cover the case of "leave-one-out" cross validation, where $k = 1$, and exactly n samples are involved, instead of 10000 as we have done. These samples are not independent, so the fluctuation will be bigger.

10. In any of the above cases, "anti cross validation", an *ad hoc* term used to denote choosing C to be equal to either A or B according to which one performs worse on the CV set, would be even more disastrous. This, however, in no way promotes the use of CV, since these are just two methods among infinitely many others.

11. The above arguments are within the framework of frequentist statistics, but they nevertheless reveal the essential role a prior must play in a theory of learning algorithms. On the other hand, if one starts from a Bayesian framework, then due to the coherence of Bayesian theory, there is no need to revert back to frequentist framework (Zhu and Rohwer, 1995a).

4 Discussions

It is well accepted that there exists prior knowledge in the practical problems, including smoothness, symmetry, positive correlation, iid samples, etc. These are indeed the implicit assumptions behind most learning algorithms. What NFL tells us is: If our algorithm is designed for such a prior, then we should say so explicitly so that a user can decide whether to use it. We cannot expect it to be also good for any other prior which we have not considered. In fact, in a sense, we should expect it to perform worse than a purely random algorithm on those other priors. Explicit treatment of smoothness priors in practical regression problems was studied in (Zhu and Rohwer, 1995b).

The power of the NFL is in a sense related to the fact that if the grand total is bound to be zero, then it is impossible to make every term positive. The apparent existence of learning rules which work well on examples without explicit requirement that these examples come from a non-uniform prior does not in any way contradict the NFL theorems, just as the apparent existence of machines doing useful work without obviously visible source of energy does not in any way contradict the principle that perpetual motion is impossible.

As early as 1775, the Parisian Academy of Science decided that they would no longer examine any invention of "perpetual motion machines", on the ground that the Law of Energy Conservation is so reliable that it will defeat any such attempt (Ord-Hume, 1977). Such a decision helped to direct human talent into the realisable effort of designing machines which utilises the energy in fuel. Should we expect the same fate for "the universally beneficial methods" in the face of NFL, and put more effort into designing methods which are good for particular priors encountered frequently in practice?

These general principles might not appear to be of obvious interest to a user, but they are of fundamental importance to a researcher. They are in fact also of fundamental importance to a user, as he must assume the responsibility of supplying the energy source, or specifying the prior.

Acknowledgement This work was partially supported by EPSRC grant GR/J17814.

References

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. Camb. Phi. Soc.*, 122:700–725.

Ord-Hume, A. W. J. G. (1977). *Perpetual Motion: The History of an Obsession.* George Allen & Unwin, London.

Wolpert, D. H. and Macready, W. G. (1995). No free lunch theorems for search. Technical Report SFI-TR-95-02-010, The Santa Fe institute.

Zhu, H. and Rohwer, R. (1995a). Bayesian invariant measurements of generalisation. `ftp://cs.aston.ac.uk/neural/zhuh/letter.ps.Z`. To appear in Neural Proc. Lett.

Zhu, H. and Rohwer, R. (1995b). Bayesian regression filters and the issue of priors. `ftp://cs.aston.ac.uk/neural/zhuh/reg_fil_prior.ps.Z`. To appear in Neural Comp. Appl.