Local Minima in Bootstrapping

Caren Marzban *
National Severe Storms Laboratory, Norman, OK 73069
Cooperative Institute for Mesoscale and Meteorological Studies, and
Department of Physics, University of Oklahoma, Norman, OK 73019

Abstract

Convergence to a global minimum is often an assumption underlying most methods for model selection, and boot-strapping is no exception. However, in many practical situations reaching a global minimum is computationally expensive. Indeed, some local minima may be much stronger attractive points than the global minimum. In this note, a simulated data set is utilized to show that boot-strapping based on the most-visited (or the "average") local minimum can yield similar results to the boot-strapping based on the global minimum.

1 Introduction

There exist a variety of methods for model selection (Bishop 1996, Ripley 1996). The difference between the various methods is usually in the way bias and variance are balanced against one another (Geman, et al.). One such balance is obtained by the method of boot-strapping (Efron and Tibshirani 1993). In its simplest form, one repeatedly trains with subsamples of the data; the average of the performance of the network on the unused subsamples is a measure of generalization performance. The number of bootstrap trials depends on the task at hand.

In each trial it is implicitly assumed that the network converges to a global minimum. Therefore, the laborious procedure of bootstrapping must be compounded by additional, computationally expensive procedures for assuring convergence to a global (or at least a very deep, local) minimum.

In this note, a simulated data set will be employed to compare and contrast the variance of the estimates due to the bootstrapping with that due to local minima. It will be shown that the variance due to bootstrapping is usually much larger than that due to the local minima, and as such it is sufficient to consider only the "average" (or the most-visited) local minimum rather than the global minimum.

*e-mail: marzban@nssl.noaa.gov and marzban@mail.nhn.ou.edu

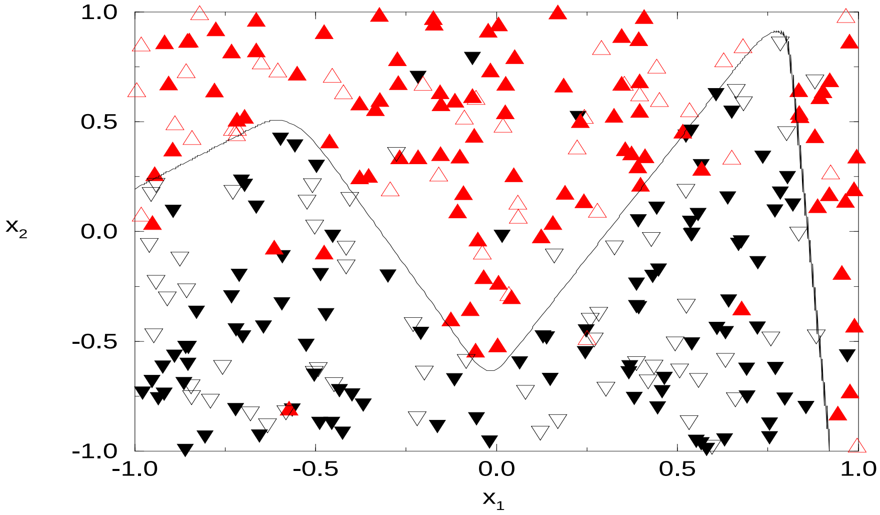

Figure 1: A subsample of data. The lower- and upper-pointing triangles represent two classes, while the "filled" and "unfilled" symbols represent the training and validation sets, respectively. The solid curve represents the underlying decision boundary.

2 Method

A 2-class classification problem is considered involving two independent, continuous variables ranging from -1 to 1. The decision boundary is shown in Figure 1. It is designed to be "learnable" by a network with one hidden layer of 4 hidden nodes, logistic activation function for both layers, minimizing cross-entropy. Noise is added to the data via a gaussian perturbation ($\sigma = 0.35$) of the boundary. Figure 1 shows one subsample of the data. The upper- and lower-pointing triangles represent the two classes, with the "filled" and "un-filled" symbols representing the training and validation sets, respectively. The size of the training set is 200 and that of the validation set is 100.

Then, a sequence of networks with $H = 0$, 2, 4, 8, and 16, hidden nodes (on one layer) is trained and validated on $S = 10$ subsamples of the simulated data. The training algorithm is conjugate gradient. When a local minimum is reached, simulated annealing is employed to attempt an escape. If a better minimum is found, then conjugate gradient is employed again, otherwise the entire training phase resumes from a new random set of initial weights. This procedure is due to Masters (1993). The total number of times that this complete reinitialization is allowed is 100, and so 100 local minima are visited. The one with the lowest value of cross-entropy over the training set is adopted as the global minimum.

First, two questions are asked regarding the bootstrap: 1) Will bootstrapping identify the global minimum of the network with $H = 4$ as the optimal model underlying the data, and 2) with what confidence? Then, the same questions are asked again, but this time with the "average" (or the most visited) local minimum of each network.

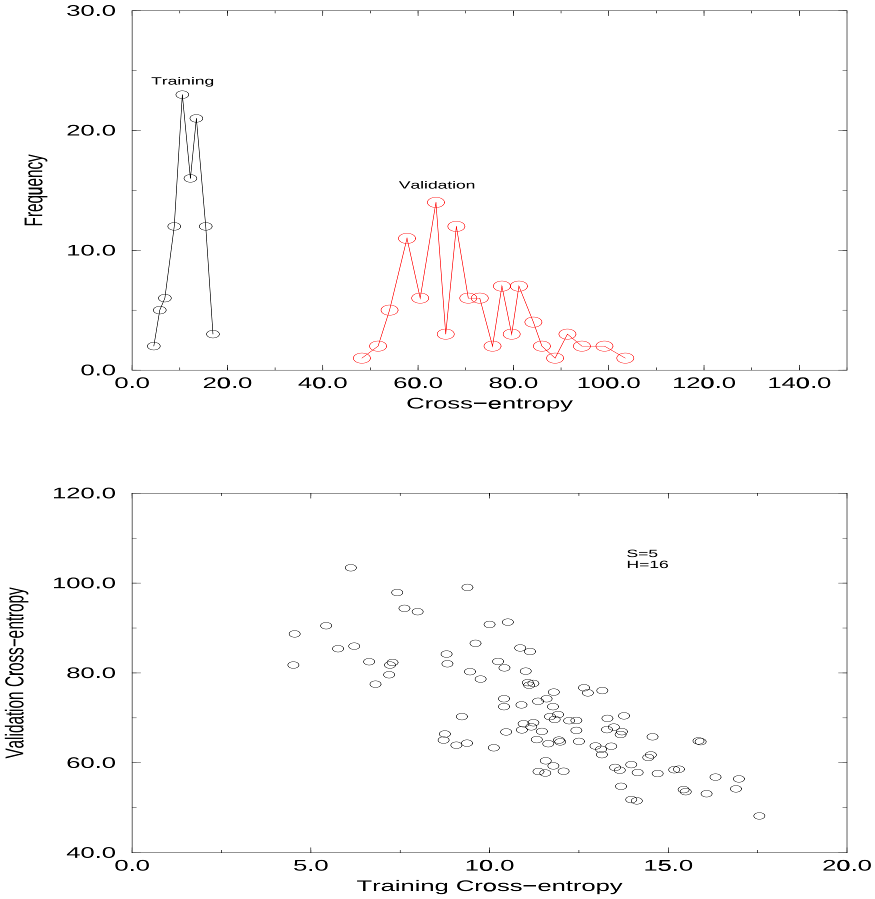

Figure 2: The distribution of 100 local minima of a network with $H = 16$ hidden nodes, for one subsample training and validation sets (top), and the correlation thereof (bottom).

3 Results

It is interesting to note that for "larger" networks, the most attractive local minimum is not a global minimum. Figure 2a shows the distribution of the local minima of a network with $H = 16$ hidden nodes trained and validated on one subsample of the data. Note that the global minimum has a cross-entropy of ~ 4, while the most visited local minimum occurs at ~ 10. The difference between the two is even more pronounced in the case of the validation set; the global minimum is at ~ 48 while the most visited local minimum is at ~ 62. One may wonder if there is any correlation between the training and the validation local minima. Figure 2b shows the (negative) correlation. Indeed, for deeper local minima tend to overfit the training set. For a "smaller" network (say, 2 hidden nodes) the correlation - not shown - is found to be positive. This is entirely expected since the optimal network is known to have $H = 4$ hidden nodes.

The answer to the first two questions is found in Figure 3a, wherein the average validation cross-entropy is plotted against the average training cross-entropy for networks with $H = 0, 2, 4, 8$, and 16 hidden nodes. The average is over the global minima of 10 subsamples, and so the error-bars are the standard deviations arising from the bootstrap. Evidently, $H = 4$ is clearly identified as the optimal number of hidden nodes. However, the uncertainty due to the bootstrap allows for the $H = 2$ and $H = 8$ networks to be treated as statistically equivalent to the $H = 4$ network.

Now, if instead of the global minimum, the average[1] of the local minima is employed, a similar pattern emerges (Figure 3b). $H = 4$ is still identifiable as the optimal model, but now there are two sources of uncertainty - due to the bootstrap ("thin" error-bars), and due to the averaging over the local minima ("thick" error-bars). Note that the former dominate the latter.

In conclusion, bootstrapping is a reliable method for model selection even if training does not necessarily converge to a global minimum. In particular, the most-visited (or the average) local minimum yields results similar to those based on the global minimum.

4 References

1. Bishop, C. M. (1996). *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, pp. 482.
2. Efron, B., and R. J. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall, London.
3. Geman, S., E. Bienenstock, and R. Doursat (1992). Neural Networks and the bias/variance dilemma. *Neural Computation*, **4** (1), 1-58.
4. Masters, T. (1993): *Practical Neural Network Recipes in C++.* Academic Press, 493 pp.
5. Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* University Press, Cambridge.

[1] In the example considered here, the average coincides with the mode of the distribution (e.g. see Figure 2a). As such, it is sufficient to consider only the former.

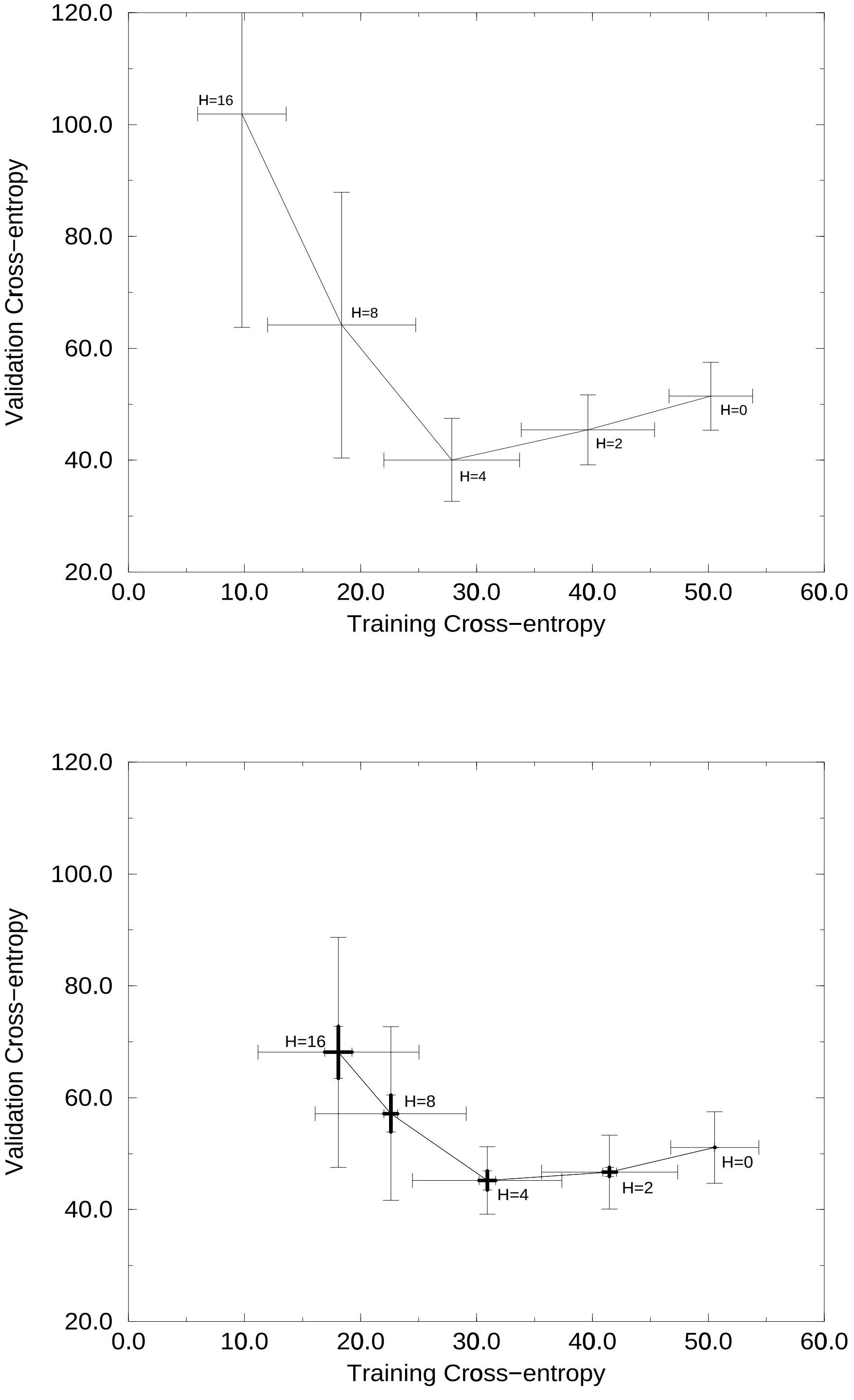

Figure 3: Illustration of model selection via bootstrapping based on the global minimum (top), and the average local minimum (bottom). The "thin" error-bars correspond to bootstrapping, while the "thick" error-bars are due to the local minima.