Bayesian Approximation Theory

David J.C. MacKay
Cavendish Laboratory, Cambridge, CB3 0HE, United Kingdom
mackay@mrao.cam.ac.uk

October 29, 2004

Abstract

Given that a learning algorithm achieves a training error $\hat{\epsilon}_{\rm M}$ on its training set, what do we expect its test error to be?

This is an inference problem ("Given A, predict B") so it must have a Bayesian answer. This note discusses the forward model and prior required to get sensible answers.

Let's take 'machine learning' to mean 'fitting a parametric or nonparametric model to data in order to make predictions'. The main activity of Bayesians in machine learning between 1990 and 2004 has been based on viewing machine learning as probabilistic modelling. We take the probabilistic model literally, as a model of the process that generated the data; then we use Bayesian inference to implement learning and prediction. This perspective has allowed researchers to (a) automate complexity control; (b) enhance the specification of the machine learning model by constructing models more in accordance with the sort of assumptions we wish to make about the data; (c) produce better predictions by using the Bayesian concept of marginalization. All these enhancements of machine learning depend on the assumption that we literally believe in the model.

An alternative attitude to the model, however, is to think of it as a black box having no relation to the process generating the data, or our beliefs about that process. The sort of question one might then ask is, 'Given that this optimized black box produced a training error of $\epsilon_{\rm t}$, what should I expect its test error to be?'

This is a question in approximation theory, and it is an inference problem. (Notice the form of the question, '**given** ..., **predict** ...'.) Thus it is a question that should be addressable in Bayesian terms. I think Wolpert (1995) was one of the first to spell out such a Bayesian approach to approximation theory.

Most of the work in approximation theory has used a sampling theory approach (Valiant, 1984; Vapnik, 1995), delivering bounds on the generalization error that are likely (in the sampling theory sense) to be true.

This note explores a Bayesian approach to the question of predicting test-set error. The motivations for a Bayesian approach are (a) to better understand

the assumptions required to make reasonable predictions of test error; (b) to obtain predictions that take all available information into account, given a set of assumptions. (One of the bizarre properties of sampling theory bounds on test error is often the bound takes an impossible value, such as an error rate greater than 100%.) For thorough prior work on Bayesian approximation theory, see Scheffer and Joachims (1998).

To make a Bayesian model to infer test error from training error, we need two things: a forward model from test error to training error; and a prior on test error.

1 A simple approach to classification error

Let's start with a simple and crude model. The learner is assumed to have available to it 2^C distinguishable parameter-settings indexed by w, where we will call C the capacity of the learner. Given a choice of w, the learner produces a binary prediction $y^{(w)}$ that depends on input variables somehow, and the teacher provides the target, t. We'll assume every learner has a true test error $\epsilon^{(w)}$ which is the probability, averaging over all inputs, that $y^{(w)}$ is not equal to t.

For simplicity, we'll assume that the errors of all learners are independent of each other and do not depend on the input. (Certainly a bad assumption, but one that allows us to focus on the interesting issues.)

Having made these assumptions, all the remains is for us to write down the forward model that maps from the true test errors $\{\epsilon^{(w)}\}_{w=1}^{2^C}$ to the training error on a training set of size N; and to define a prior on $\{\epsilon^{(w)}\}_{w=1}^{2^C}$.

The forward model is simple: all 2^C settings of w make independent errors at their own independent rates $\{\epsilon^{(w)}\}_{w=1}^{2^C}$. For each w, the number of errors r_w has a binomial distribution.

$$P(r_w \mid \epsilon^{(w)}, N) \;=\; \binom{N}{r_w} {\epsilon^{(w)}}^{r_w} (1-\epsilon^{(w)})^{N-r_w}. \tag{1}$$

The training error can be defined to be

$$\hat{\epsilon}^{(w)} \equiv \frac{r_w}{N}. \tag{2}$$

Attention will shortly focus on a couple of parameter settings, $w_\star$ and $w_{\rm M}$, which are respectively the w whose *test error* $\epsilon^\star$ is actually the smallest of all, and the w with smallest training error, which we assume is returned by a learning algorithm. But for a moment let's imagine we've enumerated all settings of w and found their training errors.

It's interesting to note that to obtain any of the standard results of approximation theory, it's going to be essential to assign a non-trivial prior $P(\{\epsilon^{(w)}\}_{w=1}^{2^C})$. If we were to assign a simple *separable* prior, for example, then our inference about the test error of any one classifier would depend only on its own training set error, and its own prior. There would be no dependence on the capacity

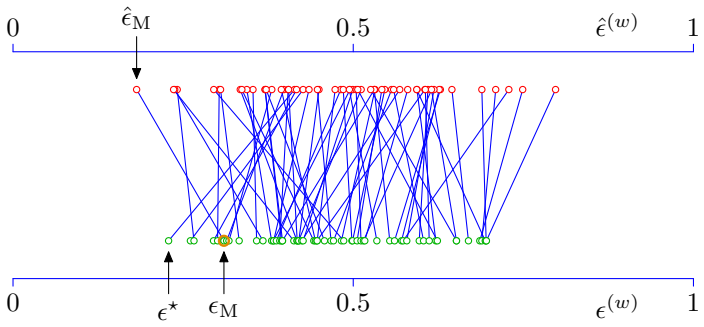


Figure 1: Schematic diagram of test errors (below) and training errors (above) for all 2^C parameter settings.

C of the learner. In contrast, experience and sampling theory both lead us to expect that the test error of the optimized parameters $w_{\rm M}$ will be greater than its training error $\hat{\epsilon}_{\rm M}$ by an amount that depends on the capacity C: bigger capacity, bigger overfitting, so bigger gap.

What prior on test errors will reproduce this reasonable intuition?

1.1 Simple model with one hyperparameter

Let's couple the $\{\epsilon^{(w)}\}_{w=1}^{2^C}$ together by introducing one hyperparameter $\epsilon^\star$, which, as defined before, is the error rate of the best classifier.

And to make things really simple, we'll assume that there exist almost 2^C other settings of w that have almost the same value of ϵ.

Then the distribution of $\hat{\epsilon}_{\rm M}$ is the distribution of the smallest of 2^C draws from a binomial distribution. Assuming that C and N are large, and making back-of-envelope approximations, we find the following forward model from $\epsilon^\star$ to $\hat{\epsilon}_{\rm M}$: given $\epsilon^\star$, the typical expected value of $\hat{\epsilon}_{\rm M}$ is the one that satisfies:

$$N D_{\rm KL}^{(2)}(\hat{\epsilon}_{\rm M}, \epsilon^\star) = C, \tag{3}$$

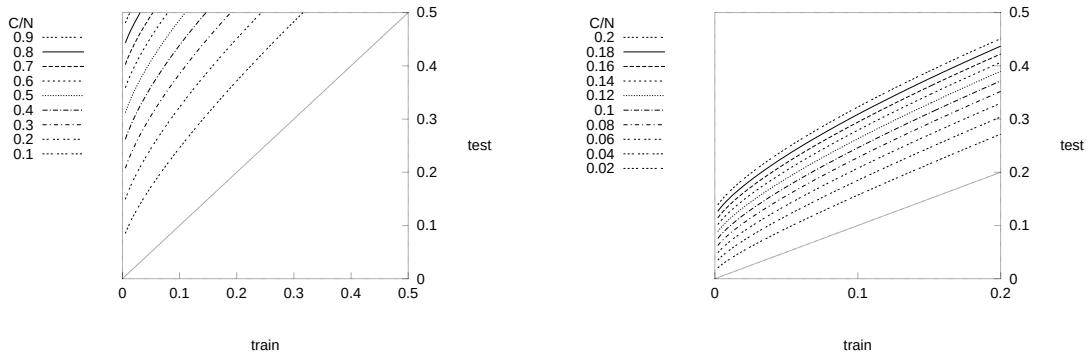
where

$$D_{\rm KL}^{(2)}(p, q) = p \log_2 \frac{p}{q} + (1-p) \log_2 \frac{(1-p)}{(1-q)}. \tag{4}$$

Here, for brevity and clarity, I haven't worked out the *distribution* of $\hat{\epsilon}_{\rm M}$ given $\epsilon^\star$, C, and N: just its typical value. Working out the distribution (which plays the role of the likelihood function) is on the to-do list.

Figure 2 shows graphs of $\epsilon^\star$ as a function of $\hat{\epsilon}_{\rm M}$ for various values of the ratio C/N.

Now, equation (3) has already appeared in the 'bounds' literature: it's in John Langford's PhD thesis, chapter 1.



(a) (b)

Figure 2: Simple model of dependence of training error on test error, as a function of the capacity-to-data ratio C/N.

1.2 More models

The above model is in a sense a worst-case model. It assumes that a huge number of models have the same test error as the best model – which sounds like good news, but it's bad news. If *lots* of models have the same test error $\epsilon^\star$, then there is a good chance that one of them will have training error *much* smaller than $\epsilon^\star$. So we have a huge *gap* between the observed best training error and the inferred test error.

A more reasonable model may be one that asserts that there are quite a lot of parameter settings with test error close to $\epsilon^\star$, but not as many as 2^C. We can introduce a curve $C(\epsilon)$ and define the number of learners as a function of test error ϵ to be

$$2^{C(\epsilon)}. \tag{5}$$

$C(\epsilon)$ must satisfy the constraint

$$\sum_\epsilon 2^{C(\epsilon)} = 2^C. \tag{6}$$

Now, each sub-population of learners with test error ϵ gets $2^{C(\epsilon)}$ chances to achieve the winning training error. The typical best training error achieved by sub-population ϵ is given by $\hat{\epsilon}_{\rm M}(\epsilon)$, which is given by the constraint (cf. equation (7))

$$N D^{(2)}_{\rm KL}(\hat{\epsilon}_{\rm M}(\epsilon), \epsilon) = C(\epsilon). \tag{7}$$

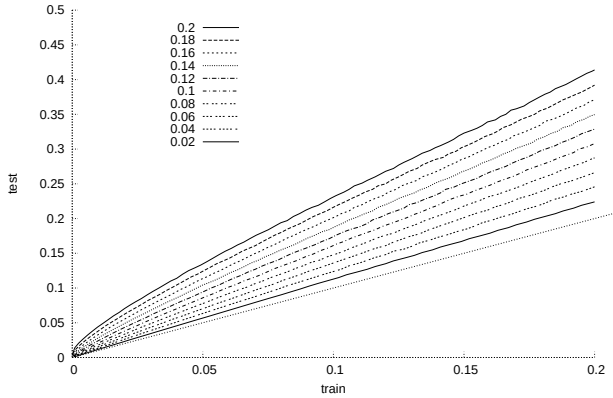


Figure 3: Second model of dependence of training error on test error, as a function of the capacity-to-data ratio C/N.

To obtain our forward model from $C(\epsilon)$ to $\hat{\epsilon}_{\rm M}$ we need to maximize $\hat{\epsilon}_{\rm M}(\epsilon)$.

To make further progress, I introduce a single-parameter family of curves $C(\epsilon)$. The one free hyperparameter is $\epsilon^\star$, the best possible test error.

$$C(\epsilon;\epsilon^\star) = CH_2\left(\frac{(\epsilon-\epsilon^\star)/2}{1/2-\epsilon^\star}\right). \tag{8}$$

I have little justification for this curve. Cases can certainly be imagined where $C(\epsilon)$ will be very different.

Now using

$$\frac{\partial}{\partial p}D_{\rm KL}(p,q) = \log\frac{p}{1-p} - \log\frac{q}{1-q} \tag{9}$$

and

$$\frac{\partial}{\partial q}D_{\rm KL}(p,q) = \frac{q-p}{q(1-q)}, \tag{10}$$

the second constraint that pins down the forward model is

$$N\frac{\epsilon-\epsilon_{\rm M}}{\epsilon(1-\epsilon)} = \frac{\partial}{\partial\epsilon}C(\epsilon;\epsilon^\star) = C\log_2\frac{1-\epsilon^\star-\epsilon}{\epsilon-\epsilon^\star}\frac{1/2}{1/2-\epsilon^\star}. \tag{11}$$

The resulting curves of typical test error versus typical training error are shown in figure 3.

Acknowledgements

I gratefully acknowledge the support of the Gatsby Charitable Foundation.

References

Scheffer, T. and Joachims, T., (1998) Estimating the expected error of empirical minimizers for model selection. Available from `http://www.informatik.hu-berlin.de/ scheffer/publications/`.

Valiant, L. G. (1984) A theory of the learnable. *Communications of the ACM* **27** (11): 1134–1142.

Vapnik, V. N. (1995) *The nature of statistical learning theory*. Springer.

Wolpert, D. H. (1995) The relationship between PAC, the statistical physics framework, the Bayesian framework, and the VC framework. In *The Mathematics of Generalization*, ed. by D. H. Wolpert. Addison-Wesley.