Bayesian Classifiers are Large Margin Hyperplanes in a Hilbert Space

Nello Cristianini
Dept of Engineering Maths
University of Bristol,
Bristol, UK
nello.cristianini@bristol.ac.uk

John Shawe-Taylor
Dept of Computer Science
RHBNC
Egham, UK
jst@dcs.rhbnc.ac.uk

Peter Sykacek
Austrian Research Institute
for Artificial Intelligence
Vienna, Austria
peter@ai.univie.ac.at

Abstract

Bayesian algorithms for Neural Networks are known to produce classifiers which are very resistent to overfitting. It is often claimed that one of the main distinctive features of Bayesian Learning Algorithms is that they don't simply output one hypothesis, but rather an entire distribution of probability over an hypothesis set: the Bayes posterior. An alternative perspective is that they output a linear combination of classifiers, whose coefficients are given by Bayes theorem. One of the concepts used to deal with thresholded convex combinations is the 'margin' of the hyperplane with respect to the training sample, which is correlated to the predictive power of the hypothesis itself.

We provide a novel theoretical analysis of such classifiers, based on Data-Dependent VC theory, proving that they can be expected to be large margin hyperplanes in a Hilbert space. We then present experimental evidence that the predictions of our model are correct, i.e. that bayesian classifers really find hypotheses which have large margin on the training examples.

This not only explains the remarkable resistance to overfitting exhibited by such classifiers, but also colocates them in the same class of other systems, like Support Vector machines and Adaboost, which have a similar performance.

Keywords: Bayesian Classifiers, Large margin hyperplanes, Hilbert space

1 INTRODUCTION

Bayesian learning algorithms for neural networks of the kind described in [3] are often claimed to have the distinctive feature of outputting an entire distribution of probability over the hypothesis space, rather than a single hypothesis. Such a distribution, the Bayes posterior, depends on the training data and on prior distribution, and is used to make predictions by averaging the predictions of all the elements of the set, in a weighted majority voting scheme.

The posterior is computed according to Bayes' rule, and such a scheme has the remarkable property that - as long as the prior is correct and the computations can be performed exactly - its expected test error is minimal. Typically, the posterior is appoximated by combining a gaussian prior and a simplified version of the likelihood (the data-dependent term, that is the term that reflects the information gleaned from the training set). Such a distribution is then sampled with a Montecarlo method, to form a committee whose composition reflects the posterior probability. The predictive integral over a posterior distribution can hence be replaced by a sum.

The classifiers obtained with this method are known to be highly resistent to overfitting. Indeed, neither the committee size nor the network size strongly affect the performance, to such an extent that it is not uncommon - in the bayesian literature - to find computations with "infinite networks" [4], [10], meaning by this the posterior over the complete (infinite) hypothesis space.

Statistical Learning Theory, on the other hand, is concerned with the problem of bounding the test error (in the worst case and with high probability) using quantities that are observable in the training set or known a priori [9].

The expressions obtained for such a bound typically depend on the training error, the sample size and the VC dimension of the classifier. Given that the number of tunable parameters gives a rough estimation of the VC dimension, the size of the network and that of the committee do matter.

A more refined, Data-Dependent, version of the theory introduced in [8], shows that it is possible to replace the VC dimension in the above mentioned bounds with a quantity which depends on the margin of the classifier on the training examples.

In this paper we provide a novel description of Bayesian classifiers which makes it possible to perform margin analysis on them, and hence to apply Data-Dependent VC theory. In particular, by viewing the posterior distribution as a linear functional in a Hilbert space, the margin can be computed and gives a bound on the generalization error via an 'effective' VC dimension which is much lower than the number of parameters.

Finally, experimental study is performed with a standard bayesian algorithm [5] on real world data, in order to test the predictions of our model. The results of the experiments confirm that the model captures the relevant features of these classifiers, and that they can indeed be regarded as large margin hyperplanes in a Hilbert space.

Margin-distribution graphs are provided for different data sets, different network sizes, committee sizes and choices of prior, always showing the same qualitative behaviour: a clear bias toward large margin on training examples.

Our plots can be directly compared with the ones presented in the inspiring paper by Shapire et al. [7], where this concept was introduced, as we have used the same datasets. In that paper, a bound on the test error as a function of the margin distribution was first obtained.

These theoretical and experimental results not only explain the remarkable resistance to evrfitting observed in bayesian algorithms, but also provide a surprising unified description of three of the most effective learning algorithms: Support Vector Machines, Adaboost and now also Bayesian classifiers.

2 BAYESIAN LEARNING THEORY

The result of Bayesian learning is a probability distribution over the (parametrized) hypothesis space, expressing the degree of belief in a specific hypothesis as approximation of the target function. Such distribution is then used to make predictions.

To start the process of bayesian learning, one must define a prior distribution $P(w)$ over the parameter space, possibily encoding some prior knowledge. After observing the data, the prior distribution is updated using Bayes' Rule:

$$P(w|D) \propto P(D|w)P(w),$$

where $P(w|D)$ is the probability of the parameters given the data D, $P(D|w)$ the probability of the data given the parameters, and $P(w)$ the prior distribution over the parameters. The posterior distribution so obtained, hence, encodes information coming from the training set (via the likelihood function $P(D|w)$) and prior knowledge.

To predict the label of a new point, bayesian classifiers integrate the predictions made by every element of the hypothesis space, weighting them with the posterior associated to each hypothesis, obtaining a distribution of probability over the set of possible labels (note that h_w is the function parametrised by w):

$$P(y|x, D) = \int_w h_w(x)p(w|D)dw$$

This predictive distribution can be used to minimize the number of misclassifications in the test set; in the 2-class case this is achieved simply by outputting the label which has received the highest vote.

3 BAYESIAN CLASSIFIERS AS LARGE MARGIN HYPERPLANES

Hence, the actual hypothesis space used by Bayesian systems is the Convex Hull of H, rather than H. The output hypothesis is a hyperplane, whose coordinates are given by the posterior.

In order to study the margin of such hyperplanes, we will introduce some simplifications in the general model. We assume that the base hypothesis space, H is formed by Boolean valued functions, and that it is sufficiently rich that all dichotomies can be implemented. Further, initially we will assume that the average prior probability over functions in a particular error shell does not depend on the number of errors.

These are the only assumptions we make, and the second will to be relaxed in a second stage. A natural choice for the evidence function in a Boolean valued hypothesis space is $e^{-k\sigma}$, where k is the number of mistakes made by the hypothesis and $\sigma > 0$ an appropriately chosen constant. The expression has the required property of giving low likelihood to the predictors which make many mistakes on the training set, and to which the usual Bayesian evidence collapses in the Boolean case. Our analysis will also suggest suitable choices for σ.

It can be interpreted with an assumption of Gaussian noise corrupting the data after they have been labelled by a target function which belongs to H, the variance of the noise depending on $1/\sigma$.

The assumption that all the dichotomies can be implemented with the same probability corresponds to an 'uninformative' prior, where no knowledge is available about the target function. In a second stage we will examine the effect of inserting some knowledge in the prior, by slightly perturbing the uninformative one towards the target hypothesis. We will see that even slightly favourable priors can give a much smaller VC dimension than the uninformative one.

3.1 THE UNINFORMATIVE PRIOR

The actual hypothesis space used by Bayesian systems, hence, is the Convex Hull of H, rather than H. The output hypothesis is a hyperplane, whose coordinates are given by the posterior.

In this section we give an expression for the margin of the composite hypothesis, as a function of a parameter related to our model of likelihood. The result is obtained in the case of a uniform prior, and for the pattern recognition case.

Let us start by stating some simple results and definitions which will be useful in the following.

Definition 3.1 *Let B_i be the* balance *of the hypothesis h_i over a given sample of size m, that is the number of successes s_i minus the number of failures f_i: $B_i = s_i - f_i$, $m = s_i + f_i$.*

Therefore $B_i = m - 2f_i$, which implies $B_i/m = 1 - 2\epsilon_i$, where $\epsilon_i = f/m$ is the empirical error of h_i.

During the next proof we will need to know the probability in the prior distribution of hypotheses in our parameter space with a fixed empirical error. Given that this information is in general not available, we will initially make the simplifying assumption that all behaviours on the training sample can be realised. This implies that the hypothesis space has VC dimension greater than or equal to the sample size m.

We make the further assumption that the prior probability of hypotheses which have error $\epsilon = k/m$ is

$$\frac{1}{2^m}\binom{m}{k} = \frac{m!}{2^m(m\epsilon)!(m-m\epsilon)!},$$

in other words that the average prior probability for functions realising different patterns of k errors is 2^{-m}. We will assume that the posterior distribution for a hypothesis which has k training errors is proportional to $e^{-\sigma k} = C^k$, where $C = e^{-\sigma}$. We are now ready to give the main result of this section.

Theorem 3.2 *Under the above assumptions the margin of the Bayes Classifier is given by*

$$1 - \frac{2C}{1+C}.$$

Proof: Let the set of training examples be $(x_1, \ldots, x_m)$ with classifications $\mathbf{y} = (y_1, \ldots, y_m) \in \{-1, 1\}^m$. Let the margin M of example i be M_i. Consider first the average margin

$$\begin{aligned} < M > &= \frac{1}{m}\sum_{i\in S} M_i = \frac{1}{m}\sum_{i\in S} y_i F(x_i) \\ &= \frac{1}{m}\sum_{i\in S} y_i \int_H a_h h(x_i) dP(h) \\ &= \frac{1}{m}\sum_{i\in S} y_i \sum_{j\in J} a_j P_j h_j(x_i) \end{aligned}$$

where h_j, $j \in J$ are representatives of each possible classification of the sample. We are denoting by P_j the prior probability of classifiers agreeing with h_j. The quantity $a_j P_j$ is the posterior probability of these classifiers, where the coefficient $a_j = Ae^{-\sigma m\epsilon_j} = AC^{m\epsilon_j}$ is the evidence, which depends only on the empirical error and the normalising constant A. By assumption, we have

$$\sum_{k \text{ error shell}} P_j = \binom{m}{k}\frac{1}{2^m}.$$

Hence,

$$\begin{aligned} < M > &= \frac{1}{m}\sum_{j\in J} a_j P_j \sum_{i\in S} y_i h_j(x_i) \\ &= \frac{1}{m}\sum_{j\in J} a_j P_j B_j \\ &= \sum_{j\in J} a_j P_j (1 - 2\epsilon_j) \\ &= 1 - 2\sum_{j\in J} a_j P_j \epsilon_j \qquad (1) \end{aligned}$$

by the observation concerning the balance B_j of h_j and the fact that the posterior distribution has been normalised, that is $1 = \int_H a_h dP(h) = \sum_{j\in J} a_j P_j$.

We now regroup the elements of the sum on the right hand side of the above equation by decomposing the hypothesis space into error shells. Hence, we can write the above sum as

$$\sum_{j\in J} a_j P_j \epsilon_j = \frac{1}{2^m}\sum_{k=0}^{m} AC^k \binom{m}{k}\frac{k}{m}. \qquad (2)$$

Solving for A and substituting, gives

$$\sum_{j\in J} a_j P_j \epsilon_j = \frac{\sum_k C^k \binom{m}{k} \frac{k}{m}}{\sum_k C^k \binom{m}{k}}$$

We can now use the equality $\sum_k C^k \binom{m}{k} = (1+C)^m$, and the observation that $\sum_k C^k \binom{m}{k} k$ can be written as $C\frac{d}{dC}\sum_k C^k \binom{m}{k} = mC(1+C)^{m-1}$ to obtain the result for the average margin.

To complete the proof we must show that the average margin is in fact the minimal margin. We will demonstrate this by showing that the margin of all points is equal. Intuitively, this follows from the symmetry of the situation, there being nothing to distinguish between different training points in the structure of the hypothesis. The formal proof relies on performing a permutation on the training points, but has had to be omitted in this shortened version. ∎

There are three relevant bounds on the generalization error in terms of the margin on the training set. We will quote all three here and then discuss their applicability in the current context. The first two appear in Schapire *et al.* [7].

Following [7], let H denote the space from which the base hypotheses are chosen (for example Neural Networks, or Decision Trees). A base hypothesis $h \in H$ is a mapping from an instance space X to {-1, +1 }.

Theorem 3.3 *Let S be a sample of m examples chosen independently at random according to D. Assume that the base hypothesis space H has VC dimension d, and let be $\delta > 0$. Then, with probability at least $1-\delta$ over the random choice of the training set S, every weighted average function $f \in C$ satisfies the following bound for all $\theta > 0$:*

$$P_D[yF(x) \le 0] \le P_S[yF(x) \le \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{d\log^2(m/d)}{\theta^2}) + log(1/\delta\right)^{1/2}\right)$$

Theorem 3.4 *Let S be a sample of m examples chosen independently at random according to D. Assume that the base hypothesis space H is finite, and let be $\delta > 0$. Then, with probability at least $1-\delta$ over the random choice of the training set S, every weighted average function $f \in C$ satisfies the following bound for all $\theta > 0$:*

$$P_D[yF(x) \le 0] \le P_S[yF(x) \le \theta] + O\left(\frac{1}{\sqrt{m}}\left(\frac{\log^2(m)\log|H|}{\theta^2}) + log(1/\delta\right)^{1/2}\right)$$

As observed by the authors, the theorem applies to *every* majority vote method, including boosting, bagging, ECOC, etc.

The third is contained in Shawe-Taylor etal [8] and involves the fat shattering dimension of the space of functions.

Theorem 3.5 *Consider a real valued function class $\mathcal{F}$ having fat shattering function bounded above by the function* afat : $\mathbb{R} \to \mathbb{N}$ *which is continuous from the right. Fix $\theta \in \mathbb{R}$. If a learner correctly classifies m independently generated examples $\mathbf{z}$ with $h = T_\theta(f) \in T_\theta(\mathcal{F})$ such that $\mathrm{er}_\mathbf{z}(h) = 0$ and $\gamma = \min|f(x_i) - \theta|$, then with confidence $1-\delta$ the expected error of h is bounded from above by*

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k\log\left(\frac{8em}{k}\right)\log(32m) + \log\left(\frac{8m}{\delta}\right)\right),$$

where $k = \mathrm{afat}(\gamma/8)$.

Since the assumption that the underlying hypothesis space can perform any classification of the training set implies that its VC dimension is at least m, we cannot expect that learning is possible in the situation described. Indeed, we have augmented the power of the hypothesis space by taking our functions from the convex hull of H which would appear to make the situation yet worse.

Hence, in order to obtain useful applications of any of the theorems we will need to consider deviations from the most general situation described above. The deviation should not have a significant impact on the margin, while reducing the expressive power of the hypotheses.

In order to apply Theorem 3.4 the number of hypotheses in the base class H must be finite. The logarithm of the number of hypotheses appears in the result. Since we have assumed that all possible classifications of the training set can be performed the number of hypotheses must be at least 2^m making the bound uninteresting. To apply this theorem we must assume that a very large proportion of the hypotheses have zero weight in the prior, while those that have significant weights in the posterior (i.e. have low empirical error) are retained. Making this assumption the bound will become significant. However, we are interested in capturing the effect of non-discrete priors, that is situations where potentially all of the base hypotheses are included, but those with high empirical error have lower prior probability.

In order to apply Theorem 3.3 the underlying hypothesis class H must be assumed to have low VC dimension in such a way that no significant impact is made on the margin. This could be achieved by removing high error functions. Note that the functions would have to be removed, in other words given prior probability 0. Hence, the bound obtained would be no better than a standard VC bound in the original space. A situation where this approach and analysis might be advantageous is where the consistent hypothesis $h_{\mathbf{y}}$ is not included in H. This will reduce the margin by approximately $a_{h_{\mathbf{y}}} 2^{-m} = (1+C)^{-m}$, since $B_{h_{\mathbf{y}}} = m$ (see equation (1)). The approximation arises from not adjusting the normalisation to take account of the missing hypothesis and is thus a very small error.

These applications are unable to take into account the prior distribution in a flexible way. In the next section we will present an application of the third approach to show how this can take advantage of a beneficial prior.

3.2 THE EFFECT OF THE PRIOR DISTRIBUTION ON THE MARGIN BOUND

We will consider the situation where the prior decays arithmetically with the error shells. In other words the prior on hypotheses with error k is multiplied by α^k for some $\alpha < 1$. We first repeat the calculations of Theorem 3.2 for this case. The sum (2) must take into account that in this case

$$\sum_{k \text{ error shell}} P_j = \alpha^k (1+\alpha)^{-m} \binom{m}{k}.$$

The factor $(1+\alpha)^m$ cancels and the factor α appears wherever C appears, that is

$$\sum_{j \in J} a_j P_j \epsilon_j = \frac{1}{(1+\alpha)^m} \sum_{k=0}^{m} A C^k \alpha^k \binom{m}{k} \frac{k}{m},$$

while

$$\frac{A}{(1+\alpha)^m} \sum_{k=0}^{m} C^k \alpha^k \binom{m}{k} = 1.$$

Hence, the margin can be computed as

$$1 - \frac{2\alpha C}{1+\alpha C}.$$

We now quote a theorem due to Gurvits [2] that bounds the fat shattering dimension of linear functionals in Banach spaces which we will need to bound the effective VC dimension.

Theorem 3.6 *[2] Consider a Banach space B of type p and the class of linear functions L of norm less than or equal to one restricted to the unit sphere. Then there is a constant D such that* $\text{fat}_L(\gamma) \leq D\gamma^{-p/(p-1)}$.

Note that for Hilbert spaces which we will consider the value of $p = 2$.

In order to apply Theorems 3.5 and 3.6 we need to bound the radius of the sphere containing the points and the norm of the linear functionals involved. Clearly, scaling by these quantities will give the margin appropriate for application of the theorem. The Hilbert space we consider is that given by the input space X with inner product

$$\langle x, y \rangle = \int_H h(x) h(y) dP(h).$$

Hence, the norm of input points is 1 and they are contained in the unit sphere as required. The linear functionals considered are those determined by the posterior distribution. The norm is given by

$$||a||^2 = \int_H a_h^2 dP(h).$$

We must compute this value for the posterior functional in the prior described above. The integral in this case is given by

$$\begin{aligned} ||a||^2 &= \sum_{j \in J} a_j^2 P_j = \sum_{k=0}^{m} A^2 C^{2k} \frac{\alpha^k}{(1+\alpha)^m} \binom{m}{k} \\ &= \frac{(1+\alpha)^m (1+\alpha C^2)^m}{(1+\alpha C)^{2m}}. \end{aligned}$$

Hnece, the bound on the fat shattering dimension becomes,

$$g(\alpha, C) := \frac{(1+\alpha)^m (1+\alpha C^2)^m}{(1+\alpha C)^{2m-2} (1-\alpha C)^2}.$$

In the rest of this section we will consider how this function behaves for various choices of C and α, showing that for careful choices of C, values of α close to 1 can give dimensions significantly lower than m, hence give good bounds on the generalization error. The analysis shows that using this approach it is possible to make use of a beneficial prior. At the same time it suggests a value of C most likely to take advantage of such a prior.

First consider the case when $\alpha = 1$. Hence,

$$g(1, C) = \frac{2^m (1+C^2)^m}{(1+C)^{2m-2} (1-C)^2}.$$

The parameter C can be chosen in the range $[0, 1)$. However, $g(1, C) \longrightarrow_{C \to 1} \infty$, while $g(1, 0) = 2^m$. Clearly, the optimal choice of C needs to be determined if the bound is to be useful. A routine calculation establishes that the value of C which minimises the expression is, $C_0 = (m - \sqrt{m-1})/(m-2)$, which gives a value of

$$g(1, C_0) = m\left(1 + \frac{1}{m-1}\right)^{m-1} \approx em.$$

This confirms that the effective VC dimension is not increased excessively provided C is chosen around $1 - 2/\sqrt{m}$. In order to study the effect of allowing α to move slightly below 1, we will perform a Taylor expansion about $\alpha = 1$.

Let $C' = \alpha C$ and the function

$$g_1(\alpha, C') := g(\alpha, C'/\alpha) = \frac{(1+\alpha)^m (1 + {C'}^2/\alpha)^m}{(1+C')^{2m-2}(1-C')^2}.$$

Note that $\left.\frac{\partial g_1(\alpha,C')}{\partial C'}\right|_{\alpha=1} = 0$, and so $\frac{\partial g(\alpha,C_0)}{\partial \alpha} = \frac{\partial g_1(\alpha,C')}{\partial \alpha} + \frac{\partial g_1(\alpha,C')}{\partial C'}\frac{dC'}{d\alpha}$. Hence,

$$\left.\frac{\partial g(\alpha, C_0)}{\partial \alpha}\right|_{\alpha=1} = \left.\frac{\partial g_1(\alpha, C')}{\partial \alpha}\right|_{\alpha=1}.$$

Differentiating gives

$$\left.\frac{\partial g_1(\alpha, C')}{\partial \alpha}\right|_{\alpha=1} = \frac{m 2^{m-1}(1 + {C'}^2)^{m-1}}{(1+C')^{2m-3}(1-C')}$$

We can now perform a Taylor series expansion of $g(\alpha, C_0)$ about $\alpha = 1$ to obtain $g(\alpha, C_0) \approx em(1 + (\alpha - 1)\sqrt{m-1})$, where we have omitted some routine calculations. Hence, the bound on the generalization error is (ignoring log factors) $\tilde{O}(1 - (1-\alpha)\sqrt{m-1})$, so that to obtain generalization error of order ϵ, we need

$$\alpha \approx 1 - \frac{1-\epsilon}{\sqrt{m-1}}.$$

Hence, for values of α very close to 1, the prior can result in very good generalization properties.

4 EXPERIMENTS

In this section we will look at some experiments where we calculated margin distributions for two data sets. We used the vehicle data and the satimage data, both taken from the StatLog [1] database. These datasets were used by [7] for a comparison of the margin distributions of Bagging and Boosting. We used satimage as provided, there are 4435 samples in the training and 2000 in the test set. The vehicle data were merged, 500 samples were used for training and 252 for testing.

4.1 EXPERIMENTAL SETUP

Both datasets are polychotomous classification problems. To arrive at a reasonable posterior probability density over weight space besides a prior we need a proper data model and likelihood term.

According to [1], the best thing we can do in the case of polychotomous classification is to use (3), the generalized logistic or softmax transformation of the output layer activations. Given distributions of hidden unit activations, which are members of the exponential family, this transformation guarantees that the network outputs may be interpreted as probabilities for classes.

$$p(C_k \mid \underline{z}) = \frac{\exp(a_k)}{\sum_{k'} \exp(a_{k'})} \qquad (3)$$

In (3) the value a_k is the value at output node k before applying softmax activation.

Having sampled a sufficient number of weights we are ready to predict. In a Bayesian framework each input value leads to a predictive distribution of network outputs. In the case of classifications, the network output is simply given by integrating over the predictive distribution. Having sampled from the posterior over weights, in our case the expectation is approximated by a sum over the weights.

The experiments were performed for both datasets with different settings. Initially we sampled 600 weights using the standard method without ARD-priors (Automatic Relevance Determination [3]). The network size was fixed to 25 hidden units for both datasets. This experiment was used to investigate the dependence of the margin distribution of the number of weights used to represent the posterior. Discarding 50 initial weights, we calculated the margin distribution of a committee consisting of the next 150 weights and compared it to the margin distribution when using all 550 remaining weights.

To assess whether the margin distribution changes while increasing the size of the network, we performed two further experiments sampling 150 weights for a network with 50 and 200 hidden units respectively, again using conventional priors without ARD. A fourth experiment should reveal the influence of an ARD-prior on the margin distribution. We sampled 150 weights for a network with 25 hidden units using an ARD-prior on the input to hidden layer weights.

Figure 1 shows plots of the resulting margin distributions for the vehicle dataset. The margin distributions for the satimage data are shown in Figure 2. Looking at the plots of the margin distributions, we see that they are different. It is interesting to investigate

[1] The data are available via the UCI machine learning repository at
`http://www.ics.uci.edu/ mlearn/MLRepository.html.`

whether these differences are significant and whether the differences in the margin distributions are correlated with the performance of the classifier on an independent test set. From theory we expect that a classifier which shows larger margins on the training data should also show a better generalization error.

For both experiments with the 200 hidden units networks we see a trend towards lower margins. This fact can be understood when remembering that the prior variance of the hidden to output weights scales inversely with the number of hidden units. Increasing the number of hidden units forces smaller hidden to output weights which leads to a smaller complexity of the network and therefore to underfitting and increased errors on the training set.

4.2 RESULTS

In order to compare the margin distribution with the generalization error, we used each classifier to predict class labels on an independent test set. The different experimental setups and the resulting generalization errors are summarized in table 1.

Table 1: Network size, information about prior distribution, committee size, and generalization error for satimage (sat) and vehicle (veh) data.

Net size	*ARD*	*Prior*	*Comm. size*	*Error sat*	*Error veh*
25	no	$\Gamma(0.05, 0.5)$	150	9.2%	15.5%
25	no	$\Gamma(0.05, 0.5)$	550	8.9%	14.7%
50	no	$\Gamma(0.05, 0.5)$	150	8.6%	13.5%
200	no	$\Gamma(0.05, 0.5)$	150	7.7%	24.2%
25	yes	$\Gamma(0.05, 0.5)$	150	9.7%	17.5%

In order to test our hypothesis that a better performance on the test set is indicated by larger margins on the training data, we will use the first experiment as reference and compare its margin distribution with the margin distributions of the second to fifth experiment.

Four one sided t-tests were used to assess whether the observed differences of means are significant. Assuming independent individual experiments, this approach suffers from the fact that the risk of having incorrectly rejected one of the hypothesis is as large as the sum of the individual significance levels. In this case we get no problem because each experiment was highly significant. In table 2 we show the generalization error, the means of the margin distributions. We expect that larger mean values of the margin distribution correspond to smaller generalization errors. Looking at the satimage experiments, we see that this is true for the large committee experiment and for the ARD-prior experiment when compared to the first experiment. For the vehicle data we see the expected correlation for both large network scenarios and for the ARD-prior experiment again comparing with the results of the first experiment.

Table 2: Generalization error and margin distributions

Satimage data		Vehicle data	
Error	Mean margin	Error	Mean margin
9.2%	0.929	15.5%	0.73
8.9%	0.932	14.7%	0.72
8.6%	0.926	13.5%	0.78
7.7%	0.898	24.2%	0.45
9.7%	0.895	17.5%	0.70

5 CONCLUSIONS

Our theoretical analysis and experimental results show that Bayesian Classifiers of the kind described in [3] can be regarded as large margin hyperplanes in a Hilbert space, and consequently can be analysed with the tools of Data-Dependent VC theory.

The non-linear mapping from the input space to the Hilbert space is given by the initial choice of network architecture, while the coordinates of the hyperplane are given by the Bayes' posterior and hence depend both on the training data and on the chosen prior.

The choice of the prior turns out to be a crucial one, since we have shown how even slightly correctly guessed priors can be translated into a much lower VC dimension of the resulting classifier (and this - coupled with high training accuracy - ensures good generalization). But even with a totally uninformative prior there is at least no harm in using these apparently overcomplex systems.

Experiments performed on real world data confirm the predictions of the model, highlighting a strong bias toward large margins in all experimental conditions and with different data sets. Their correlation with test error has also been studied.

The practical utility of VC bounds, however, does not lie in quantitative predictions of the test error (the price for their universality is often a certain looseness), but rather in providing an analytical expression of the test error which can be used to study the role of the different parameters and design choices on the final performance. Also, via the SRM principle, such bounds provide a theoretically sound indicator of performance. The results obtained in this work can be incorporated in actual learning systems, to provide for example an independent stopping criterion: the VC bound on the error could be calculated during the learning, and the training could be stopped when no significant increase in performance is observed. Also, the other choices like net size, committee size, type of prior, could be performed using as a guideline their effect on the margin.

On the theoretical side, the surprising result of this paper is to co-locate Bayesian Classifiers in the same category of other systems – namely Support Vector Machines and Adaboost – which were motivated by very different considerations but which exhibited very similar behaviours (e.g. with respect to overfitting).

A unified analysis of the three systems is now possible, which can make potentially fruitful comparisons or cross-fertilizations much easier.

Acknowledgements

Nello Cristianini is funded by EPSRC research grant number GR/L28562.

This work was supported by the Australian Research Council, and the ESPRIT Working Group in Neural and Computational Learning (NeuroCOLT Nr. 8556).

P. Sykacek is funded by the European Commission (Biomed-2 project SIESTA, grant BMH4-CT97-2040). The Austrian Research Institute for Artificial Inteligence is funded by the Austrian Federal Ministery of Science and Transport.

The authors would like to express their gratitude to the Isaac Newton Institute for Mathematical Sciences for having been invited for several stays during the Machine Learning and Neural Networks programm, organized at the institute between August and December 1997. This joint work was initiated there.

Thanks also to Chris Williams for useful discussions and to Radford Neal for making his code freely available over the Internet.

References

[1] C. M. Bishop. *Neural Networks for Pattern Recognition.* Clarendon Press, Oxford, 1995.

[2] Leonid Gurvits, A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces, In *Proceedings of Algorithmic Learning Theory, ALT-97*, and as NECI Technical Report, 1997.

[3] Radford Neal, *Bayesian Learning in Neural Networks*, Springer Verlag, 1996

[4] Radford Neal, *Priors for Infinite Networks*, Technical Report CRG-TR-94-1 (Dept. of Computer Science, University of Toronto), `http://www.cs.toronto.edu/~radford/pin.ps.Z`.

[5] `http://www.cs.toronto.edu/~radford/software-online.html`

[6] Carl Rasmussen, *Evaluation of Gaussian Processes and Other Methods for Non-Linear Regression*, PhD Thesis. `http://www.cs.toronto.edu/pub/carl/thesis.ps.gz`

[7] R. Schapire, Y. Freund, P. Bartlett, W. Sun Lee, Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods, *Proceedings of the International Conference on Machine Learning (ICML'97)*, 1997

[8] John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, Martin Anthony, Structural Risk Minimization over Data-Dependent Hierarchies, NeuroCOLT Technical Report NC-TR-96-053, 1996. (`ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports`).

[9] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995

[10] Chris Williams, *Computation with Infinite Networks*, in Advances in Neural Information Processing Systems, NIPS 96, Morgan Kaufmann, 1997.

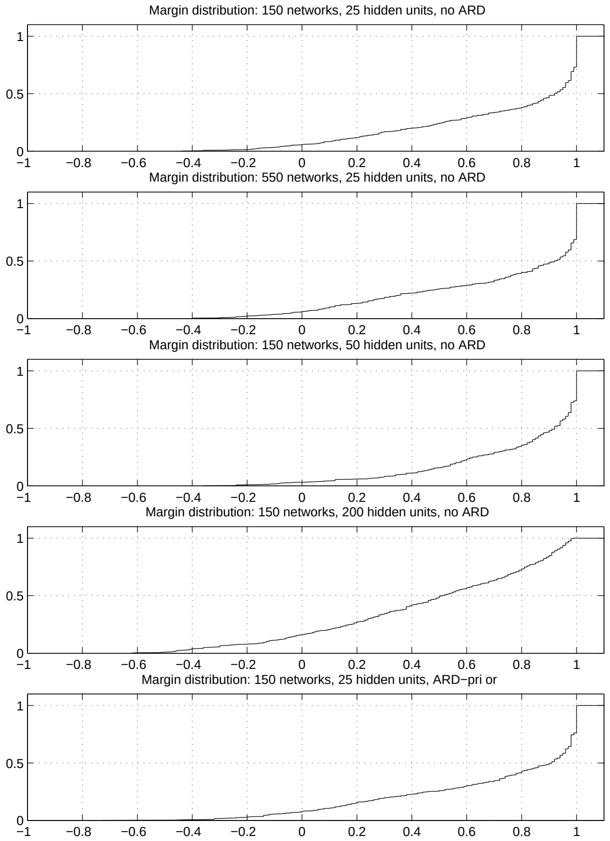

Figure 1: Plot of margin distribution of the vehicle data. The different experimental setups lead to different margin distributions. Further investigations show that these differences are highly significant. Using the first experiment as reference, the third to fifth margin distribution indicate the correct trend in the generalization error for the third to fifth classifier respectively, whereas the conclusion we would draw from the second margin distribution is misleading.

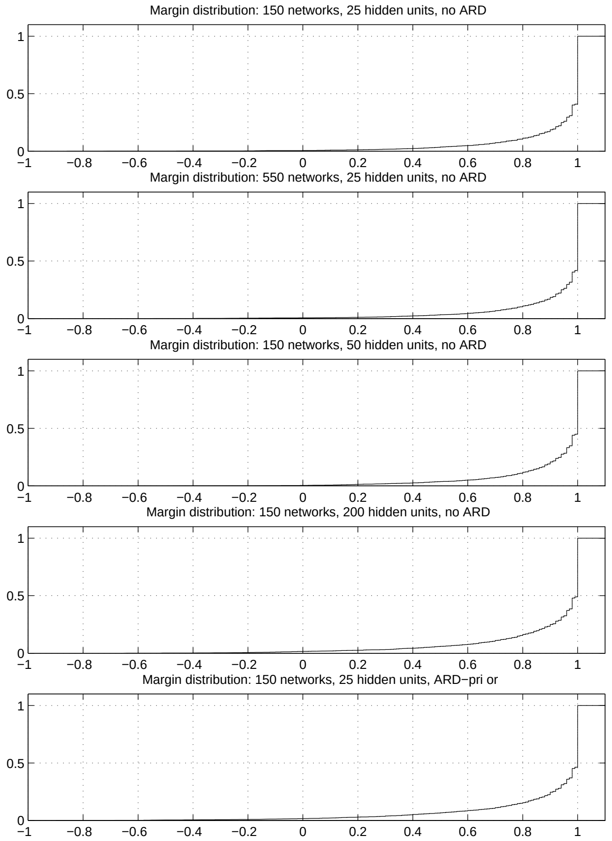

Figure 2: Plot of margin distribution of the satimage data. Also in this case we get different margin distributions. Again using the first experiment as reference, the margin distributions of these experiments allow to predict the correct trend of the generalization performance for the second and fifth experiment. The conclusion of the third and fourth margin distribution which indicates worse generalization performance compared to the first experiment is again misleading.