Bayesian Classifiers are Large Margin Hyperplanes in a Hilbert Space

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

John Shawe-Taylor
Royal Holloway, University of London
Egham, UK
j.shawe-taylor@dcs.rhbnc.ac.uk

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

Abstract

It is often claimed that one of the main distinctive features of Bayesian Learning Algorithms for neural networks 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. This can be regarded as a hyperplane in a high-dimensional feature space. 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, and hence to have low effective VC-dimension. We also present an extensive experimental study confirming this prediction. This not only explains the remarkable resistance to overfitting exhibited by such classifiers, but also co-locates them in the same class as other systems, such as Support Vector Machines and Adaboost, which have a similar performance.

1 Introduction

In recent years, new methods for training neural networks inspired by Bayesian theory have been proposed and used, mainly due to the work of MacKay and Neal [7, 8, 10]. The systems inspired by these approaches are generally known as Bayesian learning algorithms and have proven to be resistant to overfitting.

We distinguish two main approaches to Bayesian learning. In the first a hypothesis is chosen with maximum a posteriori probability, the so-called MAP estimate. Analysis of the generalization of this estimate in terms of its a posteriori probability has been made by Shawe-Taylor and Williamson [17] and McAllester [9].

In the second approach which we will study in this paper the algorithm outputs 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 the 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 approximated by combining a Gaussian prior and a simplified version of the likelihood (the data-dependent term). Such a distribution is then sampled with a Monte-Carlo method, to form a committee whose composition reflects the posterior probability.

The classifiers obtained with this method are known to be resistant 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 refer to "infinite networks" [11, 20], meaning by this networks whose number of tunable parameters is much larger than the sample size.

The thresholded linear combination of classifiers generated by the Bayesian algorithm can be regarded as a hyperplane in a high dimensional feature space. The mapping from the input to the feature space depends on the chosen hypothesis space (e.g. network architecture).

In this paper we provide a novel description of Bayesian classifiers which makes it possible to perform a margin analysis on them, and hence to apply data-dependent SRM theory [15]. 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. An analogous analysis has been performed in the case of Adaboost by Schapire *et al.*[14], whose theorems we will quote for reference. We then present an extensive experimental study substantially confirming the predictions of the model on many real world datasets.

These results not only explain the resistance to overfitting 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 (parametrised) hypothesis space, expressing the degree of belief in a specific hypothesis as an approximation of the target function. This distribution is then used to make predictions.

To start the process of Bayesian learning, one must define a prior distribution $P(\lambda)$ over the parameter space Λ associated with a set of parametrized functions $f(x, \lambda)$, possibly encoding some prior knowledge.

In the following we will denote by f_λ the hypothesis $f(x, \lambda)$. We assume throughout that f_λ are $\{-1, 1\}$ valued functions.

After observing the data D, the prior distribution is updated using Bayes' Rule:

$$P(\lambda|D) \propto P(D|\lambda)P(\lambda).$$

The posterior distribution so obtained, hence, encodes information coming from the training set (via the likelihood function $P(D|\lambda)$) 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:

$$P(y|x, D) = \int_\lambda f(x, \lambda)p(\lambda|D)dP(\lambda)$$

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.

Many practical problems exist in the implementation of such systems, and typically the procedure described above is approximated with numerical methods, by forming a committee sampled from the posterior with a Monte-Carlo simulation.

The likelihood, $P(D|\lambda)$, also needs to be approximated, and generally it is replaced by a function of the kind $e^{-\mathtt{loss}(f_\lambda))}$, meaning by this that hypotheses highly inconsistent with the training set are unlikely to have generated it, and vice-versa. The exact form taken by the likelihood, however, depends on assumptions made about the noise in the data. An introduction to this field can be found in Radford Neal's book [10]. The most important fact about Bayesian algorithms is that they turn out to be resistant to overfitting [12, 10], to the point that it is possible to use networks larger than the number of training examples, and to combine them in large committees. They are interesting not only because they work, but also because their behaviour seems to challenge the intuition of Ockham's Razor.

3 Bayesian Classifiers as Large Margin Hyperplanes

In this section we introduce a rather different view of Bayesian Classifiers, which leads to their reinterpretation as hyperplanes in a high-dimensional Hilbert space. We then study a simplified model of such classifiers, which is easier to analyse but retains all the relevant features of the general case. We wish to understand the properties of their margin, and so of their effective VC dimension. This concept was introduced by Vapnik *et al.* [19], though we use the term to mean the fat shattering dimension measured at the scale of the observed margin. Theorem 3.3 below shows that this dimension takes the place of the standard VC dimension in bounds on the generalization error in terms of the margin on the training set.

We first observe that, in the 2-class case examined so far, the predictions are actually performed by a thresholded linear combination of base hypotheses. The coefficients of the linear combinations are the posterior probabilities associated to each element of H, and the thresholding is at zero if the labels are $\{-1, +1\}$.

Hence, the actual hypothesis space used by Bayesian systems is the convex hull of H, $\mathcal{C}(H)$ rather than H, where we have

$$\mathcal{C}(H) = \left\{ F_a \middle| F_a(x) = \int_\lambda a_\lambda f(x,\lambda) dP(\lambda) \text{ where } \int_\lambda a_\lambda dP(\lambda) = 1 \right\}.$$

Hence we can view the output hypothesis is a hyperplane, whose coordinates are given by the posterior. In practice the output hypothesis is frequently estimated by a Monte-Carlo sampling of the hypothesis space using the posterior distribution. We will ignore the

effect that this has and study the behaviour of the composite hypothesis itself under various assumptions about the underlying function space H and prior $P(\lambda)$.

We first give some necessary definitions.

Definition 3.1 *Let H be a set of binary valued functions. We say that a set of points X is* shattered *by H if for all binary vectors b indexed by X, there is a function $f_b \in H$ realising b on X. The* Vapnik-Chervonenkis (VC) dimension VCdim(H) *of the set H the size of the largest shattered set, if this is finite or infinity otherwise.*

Definition 3.2 *Let $\mathcal{H}$ be a set of real valued functions. We say that a set of points X is* γ-shattered *by $\mathcal{H}$ if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b indexed by X, there is a function $f_b \in \mathcal{H}$ satisfying*

$$f_b(x) \begin{cases} \geq r_x + \gamma & \text{if } b_x = 1 \\ \leq r_x - \gamma & \text{otherwise.} \end{cases}$$

The fat shattering dimension $\text{fat}_{\mathcal{H}}$ *of the set $\mathcal{H}$ is a function from the positive real numbers to the integers which maps a value γ to the size of the largest γ-shattered set, if this is finite or infinity otherwise.*

We will make critical use of the following result contained in Shawe-Taylor *et al* [15] which involves the fat shattering dimension of the space of functions. Note that T_θ denotes the classification function obtained by thresholding at θ.

Theorem 3.3 *Consider a real valued function class $\mathcal{H}$ having fat shattering function bounded above by the function* $\text{afat} : \Re \to \mathcal{N}$ *which is continuous from the right. Fix $\theta \in \Re$. Then with probability at least $1 - \delta$ a learner who correctly classifies ℓ independently generated examples $\mathbf{z}$ with $h = T_\theta(f) \in T_\theta(\mathcal{H})$ with zero training error and $\gamma = \min |f(\mathbf{x}_i) - \theta|$ will have error of h bounded from above by*

$$\epsilon(m, k, \delta) = \frac{2}{\ell}\left(k \log_2\left(\frac{8e\ell}{k}\right)\log_2(32\ell) + \log_2\left(\frac{8\ell}{\delta}\right)\right),$$

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

More recently results bounding the generalization in terms of other more robust measures of the distribution of margin values have been obtained [16], but for the analysis presented in this paper the above theorem will be adequate.

Note how the fat shattering dimension at scale $\gamma/8$ plays the role of the VC dimension in this bound. This result motivates the use of

the term effective VC dimension for this value. In order to make use of this theorem, we must have a bound on the fat shattering dimension and then calculate the margin of the classifier. We begin by considering bounds on the fat shattering dimension. The first bound on the fat shattering dimension of bounded linear functions in a finite dimensional space was obtained by Shawe-Taylor *et al.* [15]. Gurvits [6] generalised this to infinite dimensional Banach spaces. We will quote an improved version of this bound (slightly adapted for an arbitrary bound on the linear operators) which is contained in [2].

Theorem 3.4 *[2] Consider a Hilbert space and the class of linear functions L of norm less than or equal to B restricted to the sphere of radius R about the origin. Then the fat shattering dimension of L can be bounded by*

$$\mathrm{fat}_L(\gamma) \leq \left(\frac{BR}{\gamma}\right)^2.$$

In order to apply Theorems 3.3 and 3.4 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 functions

$$\mathcal{H} = \left\{ \mathbf{z} : \Lambda \to \Re \,\middle|\, \text{such that } \int_{\lambda\in\Lambda} \mathbf{z}(\lambda)^2 dP(\lambda) < \infty \right\}$$

with the inner product

$$\langle \mathbf{z}_1 \cdot \mathbf{z}_2 \rangle = \int_{\lambda\in\Lambda} \mathbf{z}_1(\lambda)\mathbf{z}_2(\lambda) dP(\lambda).$$

There is a natural embedding of the input space X onto the unit sphere of $\mathcal{H}$ given by $\mathbf{x} \mapsto f(\mathbf{x}, \cdot)$, since

$$\int_{\lambda\in\Lambda} f(\mathbf{x},\lambda)^2 dP(\lambda) = \int_{\lambda\in\Lambda} dP(\lambda) = 1.$$

Hence, the norm of input points is 1 and they are contained in the unit sphere as required.

The linear functions considered are those determined by the posterior distribution. The norms are given by

$$\|a\|^2 = \int_\lambda a_\lambda^2 dP(\lambda).$$

Hence,

$$\mathrm{fat}_{\mathcal{C}_B(H)}(\gamma) = \left(\frac{B}{\gamma}\right)^2, \text{ where } \mathcal{C}_B(H) = \left\{ F_a \in \mathcal{C}(H) \,\middle|\, \|a\|^2 \leq B \right\}.$$

Next we consider the margin γ. 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 sufficiently rich that all dichotomies can be implemented. Further, initially we will assume that the average prior probability over functions in each error shell does not depend on the number of errors.

These are the only assumptions we make, and the second will be relaxed in a later analysis. A natural choice for the evidence function in a $\{-1, 1\}$ valued hypothesis space is $e^{-r\sigma}$, for a hypothesis that makes r mistakes. which 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. The quantity σ is usually related to the kind of noise assumed to affect the data.

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 effective VC dimension than the uninformative one.

3.1 An Uninformative Prior

The actual hypothesis space used by Bayesian systems, hence, is the convex hull $\mathcal{C}(H)$, 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 for the pattern recognition case.

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

Definition 3.5 *Let $\mathtt{s}_\lambda$ be the number of points whose labeling is incorrectly predicted by the hypothesis f_λ. We define the* balance *of the hypothesis f_λ over a given sample as $B_\lambda = \ell - 2\mathtt{s}_\lambda$, where ℓ is the sample size. Hypotheses having the same value of $\mathtt{s}$ are said to form an* error shell.

Note that $B_\lambda/\ell = 1 - 2\epsilon_\lambda$, where $\epsilon_\lambda = s_\lambda/\ell$ is the empirical error of f_λ.

Formally our initial assumption can now be stated.

Assumption 3.6 The Uninformative Prior *The prior probability of hypotheses which have empirical error $\epsilon = r/\ell$ is*

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

in other words that the average prior probability over the equivalence classes of functions realising different patterns of r errors is $2^{-\ell}$.

The posterior distribution for a hypothesis which has r training errors is proportional to $e^{-\sigma r} = C^r$, where $C = e^{-\sigma}$. We are now ready to give the main result of this section.

Theorem 3.7 *If Assumption 3.6 holds, then the margin of the Bayes Classifier $F(x) \in \mathcal{C}(H)$ is given by*

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

Proof: Let the set of training examples be $(\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$ with classifications $\mathbf{y} = (y_1, \ldots, y_\ell) \in \{-1, 1\}^\ell$ and let the margin M of example i be $M_i = y_i F(\mathbf{x}_i)$. Consider first the average margin

$$\begin{aligned} < M > &= \frac{1}{\ell}\sum_{i\in S} M_i = \frac{1}{\ell}\sum_{i\in S} y_i F(\mathbf{x}_i) = \frac{1}{\ell}\sum_{i\in S} y_i \int_{\lambda\in\Lambda} a_h h(\mathbf{x}_i) dP(h) \\ &= \frac{1}{\ell}\sum_{i\in S} y_i \sum_{j\in J} a_j P_j f_j(\mathbf{x}_i), \end{aligned}$$

where f_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 f_j. The quantity $a_j P_j$ is the posterior probability of these classifiers, where the coefficient $a_j = Ae^{-\sigma\ell\epsilon_j} = AC^{\ell\epsilon_j}$ is the evidence, which depends only on the empirical error and the normalising constant A. By assumption, we have

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

Hence,

$$\begin{aligned} < M > &= \frac{1}{\ell}\sum_{j\in J} a_j P_j \sum_{i\in S} y_i f_j(\mathbf{x}_i), = \frac{1}{\ell}\sum_{j\in J} a_j P_j B_j \qquad (1) \\ &= \sum_{j\in J} a_j P_j (1 - 2\epsilon_j) = 1 - 2\sum_{j\in J} a_j P_j \epsilon_j, \end{aligned}$$

by the observation concerning the balance B_j of f_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 (subsets of H formed by hypotheses with the same error r).

Hence, we can write the above sum as

$$\sum_{j \in J} a_j P_j \epsilon_j = \frac{1}{2^\ell} \sum_{r=0}^{\ell} A C^r \binom{\ell}{r} \frac{r}{\ell}. \tag{2}$$

Solving for A and substituting, gives

$$\sum_{j \in J} a_j P_j \epsilon_j = \frac{\sum_k C^r \binom{\ell}{r} \frac{r}{\ell}}{\sum_r C^r \binom{\ell}{r}}$$

We can now use the equality $\sum_r C^r \binom{\ell}{r} = (1+C)^\ell$, and the observation that $\sum_r C^r \binom{\ell}{r} r$ can be written as $C \frac{d}{dC} \sum_r C^r \binom{\ell}{r} = \ell C (1+C)^{\ell-1}$ 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.

More formally, note that for every output sequence

$$\mathbf{z} = (z_1, \ldots, z_\ell) \in \{-1, 1\}^\ell,$$

we can realise the mapping $\mathbf{x}_i \mapsto z_i, i = 1, \ldots, \ell$, with a function $f_{\mathbf{z}} \in H$.

Let $s(\mathbf{z})$ be the sequence obtained by swapping the i-th and j-th entries in the sequence $\mathbf{z}$ and swapping their signs if the i-th and j-th inputs have opposite classifications according to the training sequence $\mathbf{y}$. Note that s is a bijection of the set of all sequences onto itself. Note also that if a_h is the posterior distribution over the function class H, $a_{f_{\mathbf{z}}} = a_{f_{s(\mathbf{z})}}$, since the number of errors of the two functions is the same – $f_{\mathbf{z}}$ is correct on input i precisely when $f_{s(\mathbf{z})}$ is correct on j, that is

$$y_i f_{\mathbf{z}}(\mathbf{x}_i) = y_j f_{s(\mathbf{z})}(\mathbf{x}_j).$$

Now consider the Bayesian posterior function

$$F(x) = \frac{1}{2^\ell} \sum_{\mathbf{z}} a_{f_{\mathbf{z}}} f_{\mathbf{z}}(x).$$

The margin of this function on the point $\mathbf{x}_i$ is

$$y_i F(\mathbf{x}_i) = \frac{1}{2^\ell} \sum_{\mathbf{z}} a_{f_{\mathbf{z}}} y_i f_{\mathbf{z}}(\mathbf{x}_i) = \frac{1}{2^\ell} \sum_{\mathbf{z}} a_{f_{\mathbf{z}}} y_i f_{s(\mathbf{z})}(\mathbf{x}_i),$$

since s is a bijection and weights are unchanged. Hence,

$$y_i F(\mathbf{x}_i) = \frac{1}{2^\ell} \sum_{\mathbf{z}} a_{f_{\mathbf{z}}} y_j f_{\mathbf{z}}(\mathbf{x}_j) = y_j F(\mathbf{x}_j)$$

and the margins of the points i and j are equal. Since, i and j are arbitrary all margins are equal and the result is proved. ■

Since the assumption that the underlying hypothesis space can perform any classification of the training set implies that its VC dimension is at least ℓ, 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.

Nonetheless Theorem 3.7 shows that the margin of the Bayes classifier is indeed large under the assumptions we have made, provided a suitable choice of the parameter C is made. A calculation of the effective VC dimension in this case will be made later, though it is too large for any bound on the generalization error to be made. We must make assumptions about the prior in order to be able to learn.

3.2 Towards a Benign Prior

Before proceeding to consider the effect of the prior on the effective VC bound, we will mention two other theorems that might be useful for bounding the generalization error in terms of the margin. We will, however, argue that they are unable to take account of our type of prior that assigns different probabilities to hypotheses. We will quote the theorems from Schapire *et al.* [14], though they appear in a more general form in [1].

Following [14], let H denote the space from which the base hypotheses are chosen (for example Neural Networks, or Decision Trees). A base hypothesis $f \in H$ is a mapping from an instance space X to $\{-1, +1\}$.

Theorem 3.8 *Let S be a sample of ℓ 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 $F \in \mathcal{C}(H)$ satisfies the following bound for all $\theta > 0$:

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

Theorem 3.9 *Let S be a sample of ℓ 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 $F \in \mathcal{C}(H)$ satisfies the following bound for all $\theta > 0$:*

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

As observed by the authors, the theorem applies to *every* majority vote method, including boosting, bagging, ECOC, etc. 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 3.9 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^ℓ 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 3.8 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 $f_{\mathbf{y}}$ is not included in H. This will reduce the margin by approximately

$$a_{f_{\mathbf{y}}} 2^{-\ell} = (1+C)^{-\ell},$$

since $B_{f_{\mathbf{y}}} = \ell$ (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. Note that the minimal margin is still equal to the average in this case since all points are equally affected. Applying a non-agnostic version of the second theorem assuming that the underlying function class had VC dimension k would now give a significantly better bound on generalization than a standard agnostic bound for the best hypothesis in the base class which would not have zero error.

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

3.3 The Effect of a Benign Prior

We will consider the situation where the prior decays arithmetically with the error shells. In other words the prior on hypotheses with error r is multiplied by ρ^r for some $\rho < 1$. More formally we state this as follows.

Assumption 3.10 The Benign Prior *The prior probability of hypotheses which have empirical error $\epsilon = r/\ell$ is*

$$\sum_{r \text{ error shell}} P(\lambda) = A\rho^r \binom{\ell}{r} = \frac{\rho^r}{(1+\rho)^{-\ell}} \binom{\ell}{r}.$$

We first repeat the calculations of Theorem 3.7 for this case. The sum (2) must take into account the new prior. The factor $(1+\rho)^\ell$ is common and the factor ρ appears wherever C appears, that is

$$\sum_{j \in J} a_j P_j \epsilon_j = \frac{1}{(1+\rho)^\ell} \sum_{r=0}^{\ell} AC^r \rho^r \binom{\ell}{r} \frac{r}{\ell},$$

while

$$\frac{A}{(1+\rho)^\ell} \sum_{r=0}^{\ell} C^r \rho^r \binom{\ell}{r} = 1.$$

Hence, $(1+\rho)^\ell$ cancels, and since the argument that all margins are equal is not affected by the prior, we can deduce the following generalization of Theorem 3.7.

Theorem 3.11 *If Assumption 3.10 holds, then the margin of the Bayes Classifier $F(x) \in \mathcal{C}(H)$ is given by*

$$\gamma_{\rho,C} = 1 - \frac{2\rho C}{1+\rho C}.$$

We must further compute the value of $\|a\|$ for the posterior functional in the prior described above. The integral in this case is given by

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

We can now combine this value with the margin computed above to give the value of the fat shattering dimension from Theorem 3.4 at the appropriate scale. This bound on the effective VC dimension becomes,

$$g(\rho, C) := \frac{\|a\|^2}{\gamma_{\rho,C}^2} = \frac{(1+\rho)^\ell (1+\rho C^2)^\ell}{(1+\rho C)^{2\ell-2}(1-\rho C)^2},$$

where to keep the subsequent formulae simple we have ignored the factor of 64 arising for the scale $\gamma/8$ in Theorem 3.3.

In the rest of this section we will consider how this function behaves for various choices of C and ρ, showing that for careful choices of C, and values of ρ close to 1 can give dimensions significantly lower than ℓ, hence give good bounds on the generalization error. The analysis shows that using this approach it is possible to understand the effect of a benign 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 $\rho = 1$, that is the uninformative prior considered in Section 3.1. Hence,

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

The parameter C can be chosen in the range $[0, 1)$. However,

$$g(1, C) \longrightarrow_{C\to 1} \infty,$$

while $g(1, 0) = 2^\ell$. 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, $C_0 = (\ell - \sqrt{\ell - 1})/(\ell - 2)$, which gives a value of

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

This confirms that the effective VC dimension is not increased excessively provided C is chosen around $1 - 2/\sqrt{\ell}$ (i.e. $\sigma \approx 2/\sqrt{\ell}$), though

of course the bound on generalization error is trivial in this case. The analysis so far can be viewed as a 'sanity check', demonstrating that despite significantly increasing the computational power of the hypothesis class (by moving to $\mathcal{C}(H)$), the increase in the effective VC dimension has been very slight. In order to see how the prior can produce a non-trivial bound, we will study the effect of allowing ρ to move slightly below 1. We will perform a Taylor expansion about $\rho = 1$.

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

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

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

Hence,

$$\frac{\partial g(\rho, C_0)}{\partial \rho}\Big|_{\rho=1} = \frac{\partial g_1(\rho, C')}{\partial \rho}\Big|_{\rho=1}.$$

Differentiating gives

$$\frac{\partial g_1(\rho, C')}{\partial \rho}\Big|_{\rho=1} = \frac{\ell 2^{\ell-1}(1+C'^2)^{\ell-1}}{(1+C')^{2\ell-3}(1-C')}$$

We can now perform a Taylor series expansion of $g(\rho, C_0)$ about $\rho = 1$ to obtain $g(\rho, C_0) \approx e\ell(1 + (\rho - 1)\sqrt{\ell - 1})$, where we have omitted some routine calculations.

Hence, the bound on the generalization error is (ignoring log factors)

$$\tilde{O}(1 - (1-\rho)\sqrt{\ell - 1}),$$

so that to obtain generalization error of order ϵ, we need

$$\rho \approx 1 - \frac{1 - \epsilon/(e \log \ell)}{\sqrt{\ell - 1}}.$$

Hence, for values of ρ very close to 1, the prior can result in improved generalization properties. Note that the value of C used in the calculations is unchanged so that we can take advantage of the prior without any fine tuning of the system. We simply observe the margin, and the value of $\|a\|$ on the Monte-Carlo generated set of hypotheses, to recover a bound on the effective VC dimension and hence an estimate of the generalization error.

One possible criticism of the benign prior of Assumption 3.10 is that in many learning systems, natural priors have the property that a function and its complement have equal probability. The assumption of arithmetically decreasing prior with increasing error shell is not

consistent with this. This problem can, however, be overcome by considering a prior that emphasises very bad hypotheses in the same way as very good hypotheses, those having low correlation (and anti-correlation) with the target being the low probability ones:

$$\sum_{r \text{ error shell}} P(\lambda) = A(\rho^r + \rho^{\ell-r}) \binom{\ell}{r} = \frac{\rho^r + \rho^{\ell-r}}{2(1+\rho)^{-\ell}} \binom{\ell}{r}.$$

The effect of this on the margin and norm is slight. For example, the margin becomes

$$\begin{aligned} \gamma &= 1 - \frac{2\rho C(1+\rho C)^{\ell-1} + C(\rho + C)^{\ell-1}}{(1+\rho C)^{\ell} + (\rho + C)^{\ell}} \\ &\approx 1 - \frac{2\rho C}{1+\rho C}. \end{aligned}$$

provided the following quantity is significantly smaller than $2\rho \approx 2$,

$$\begin{aligned} \left(\frac{\rho + C}{1+\rho C}\right)^{\ell-1} &= \left(1 - \frac{(1-\rho)(1-C)}{1+\rho C}\right)^{\ell-1} \\ &\approx \left(1 - \frac{1}{\ell}\right)^{\ell} \\ &\approx e^{-1} \end{aligned}$$

for the values of ρ and C considered. We have omitted a full derivation as the formulae become rather unwieldy.

4 Experiments

In the previous sections we have presented a slightly simplified model of a Bayesian classifier, which predicted that the margin of the training points can be expected to be large. When real data are used, however, the assumptions of the model could be only partially satisfied, so also its predictions could be affected. In this section we present an extensive experimental study on real world data, showing that indeed the committees of neural networks produced by Bayesian classifiers do generate large margin hypotheses.

Even if not all data points have margin one, which was the idealized situation of our model, we still can see that this is nearly the case: namely that the distribution of the margins is clearly biased in the sense of a large margin. This is enough for us to use margin distribution theorems such as [16], and bound the effective VC dimension of such systems, so explaining their seemingly paradoxical behaviour with respect to overfitting.

4.1 Data

The datasets were chosen to allow comparisons with [14] and to cover a range of different problems. We used vehicle data, satimage data, Wisconsin breast cancer, lung cancer data, John Hopkins University ionosphere data, balance scale weight and distance data and the wine recognition database, all taken from the StatLog [1] database. 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. The other data were split into two equal sized sets, which were both used as training and independent test sets respectively. The samples with missing values present in the lung cancer data were removed. To cover also categoric problems, we used the titanic dataset, which is provided via the Delve project [2]. Here we used the largest number of training data suggested, which is 500 samples and the test set as provided, which contains 1701 samples. The data has 3 categoric inputs: gender, age (adult/child) and the class of passage (first to third class or crew member). We used ordered coding of the information about the passenger class, which seems to be a reasonable choice. The last data is the pima diabetes dataset as provided by B.D. Ripley[3]. A detailed description can be found in [13].

4.2 Experimental setup

We performed two types of experiments: For satimage, vehicle, titanic and the pima diabetes data, we looked at the performance resulting from five different settings of the classifier and learning procedure. All experiments were performed using R. Neals hybrid Monte Carlo sampling algorithm. Initially we sampled 600 weights using the standard method without automatic relevance determination (ARD)-priors. The network size was fixed to 25 hidden units. 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

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

[2]The Delve home-page can be reached at http://www.cs.utoronto.ca/ delve/.

[3]The data can be obtained via: http://www.stats.ox.ac.uk/ ripley/.

respectively, again using conventional priors without ARD. A fifth 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.

The term automatic relevance determination (ARD) -prior refers to an additional level in the hierarchical prior specification. In R. Neals hybrid Monte Carlo implementation, ARD-priors can be used on groups of weights connecting one unit (typically an input or a hidden unit) to the next layer. As mentioned, we used ARD on inputs only. The idea of this concept is to allow "soft" feature selection. To achieve this, each group of weights from one input to the hidden layer has it's own Gaussian prior governed by a hyper parameter (the variance of the Gaussian). If most weights from one input tend to be small, the posterior of the associated hyperparameter favors small variances. Hence the weights are actively suppressed and the input is "turned off".

Figure 1 shows plots of the resulting margin distributions for the satimage, vehicle, titanic and pima data. Looking at the plots of the margin distributions for each experiment, we see a trend towards lower margins when increasing the number of hidden units. 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 increased errors on the training set.

The setting for the experiments performed with the remaining datasets (balance scale, breast cancer, ionosphere, lung cancer and the wine recognition data) was somewhat simpler. Our first experiment gave some evidence that significant differences of the observed performance were mainly due to differences caused by the ARD-prior. Therefore, we restricted our second comparison to results obtained with a two layer neural network with fifteen hidden units, once with conventional (non ARD) priors and once with ARD-priors. Both equal sized datasets were used as training and test set respectively. In figure 2 we see the margin distributions on the training data when fusing the margins of both runs.

4.3 Discussion of margin distributions

In order to compare the margin distribution with the generalization error, we used each classifier to predict class labels on an independent test set. In table 1, we summarize the different experimental settings and the results obtained on all datasets. In the order of their

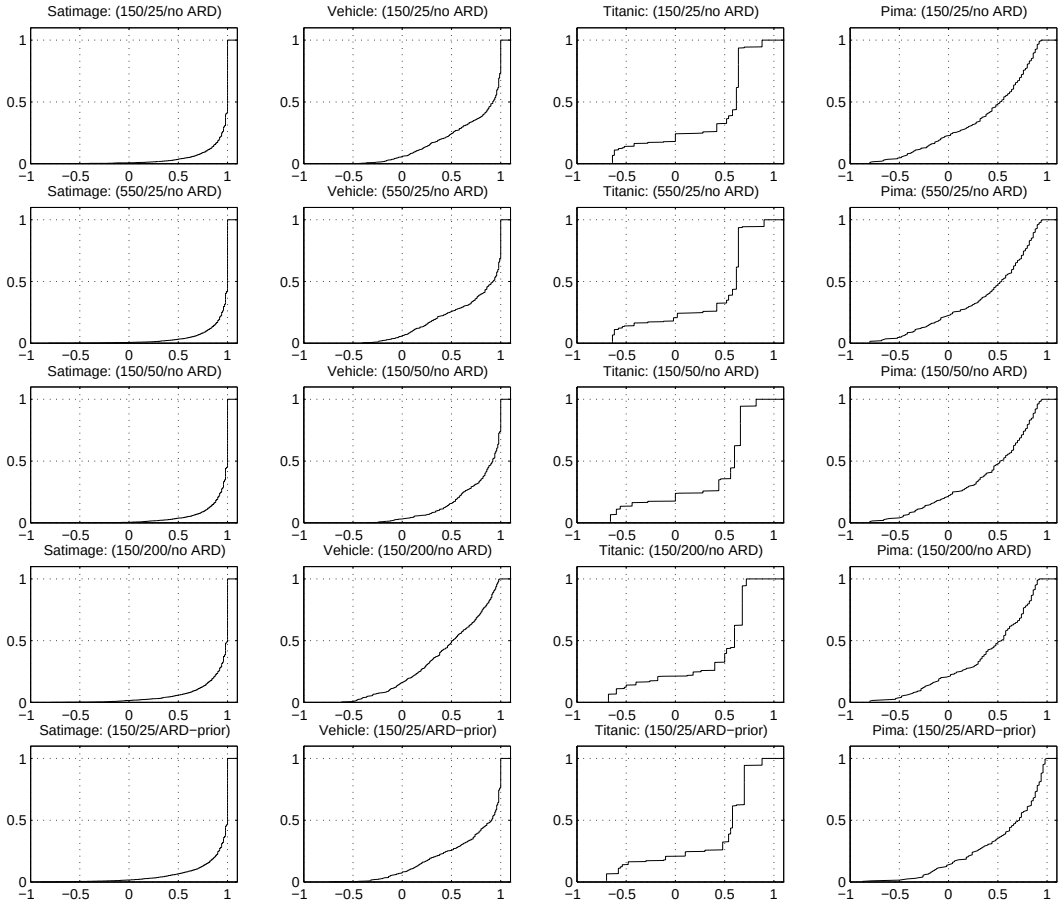

Figure 1: Plots showing margin distributions of several standard data sets using different network architectures and settings for the sampling algorithm. The different experiments show that the margin distributions are correlated with the generalization performance observed on an independent test set. A remarkable property which can be observed here, is that networks with a *larger* number of hidden units tend to *less* complex hypothesis classes. Our explanation for this effect is the scaling for the variance of the priors over output weights used in R. Neals sampling algorithm, which is inverse to the number of hidden units.

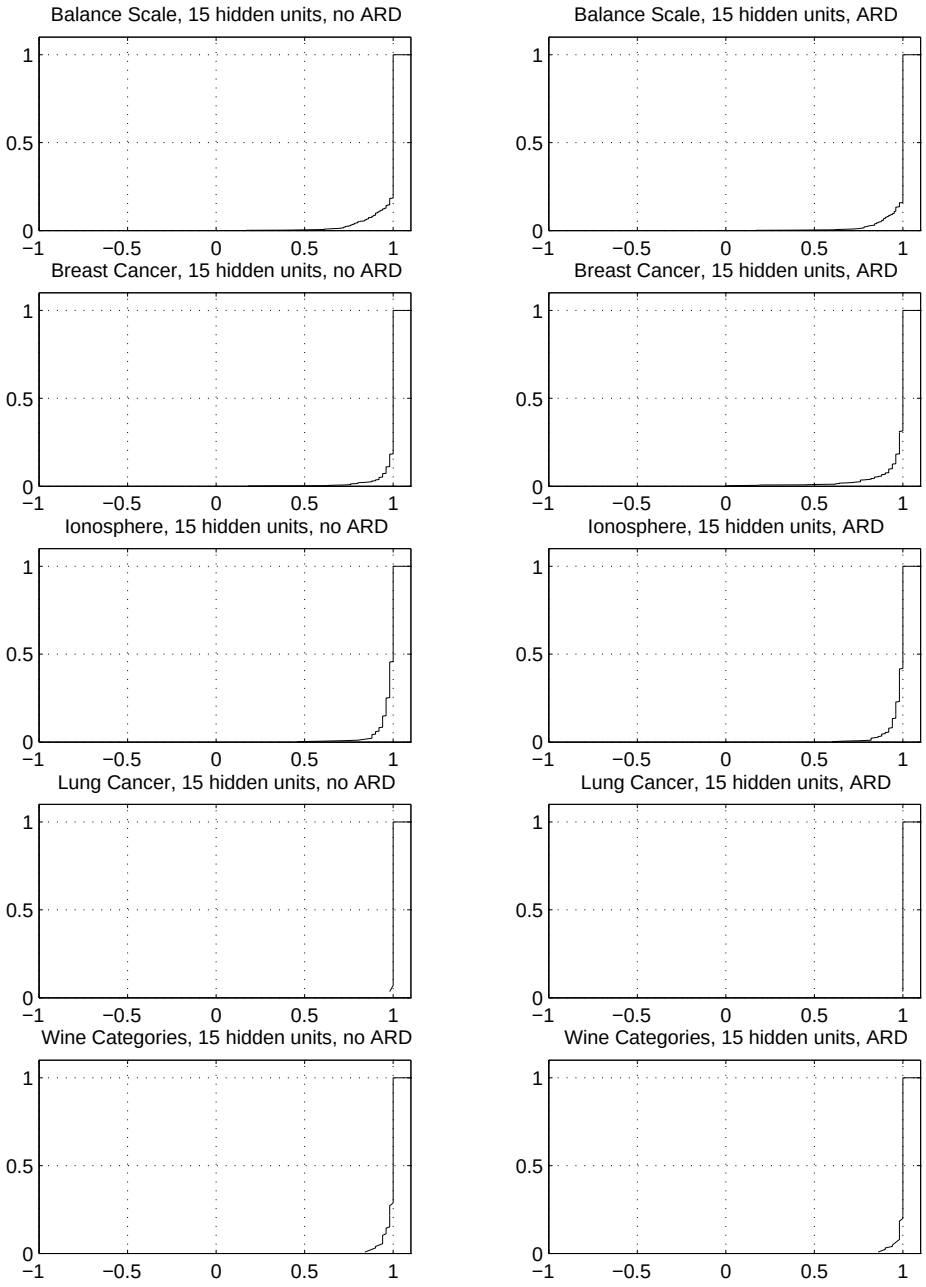

Figure 2: Plots showing margin distributions of five standard data sets using a neural network with fifteen hidden units, once without and once with ARD-prior.

Table 1: Summary of Experiments

Dataset	Network size	Committee size	ARD-prior	Gen. error	Mean of margins
Satimage	25	150	no	9.2%	0.929
	25	550	no	9.0%	0.932
	50	150	no	8.6%	0.926
	200	150	no	9.7%	0.895
	25	150	yes	8.6%	0.899
Vehicle	25	150	no	15.5%	0.727
	25	550	no	14.7%	0.720
	50	150	no	13.5%	0.782
	200	150	no	17.5%	0.698
	25	150	yes	23.0%	0.458
Titanic	25	150	no	22.9%	0.366
	25	550	no	22.9%	0.367
	50	150	no	22.0%	0.3637
	200	150	no	23.1%	0.377
	25	150	yes	22.8%	0.57
Pima	25	150	no	20.2%	0.392
	25	550	no	19.9%	0.397
	50	150	no	21.1%	0.402
	200	150	no	25.6%	0.557
	25	150	yes	20.2%	0.402
balance scale	15	150	no	13.6%	0.975
	15	150	yes	16.3%	0.982
breast cancer	15	150	no	3.0%	0.987
	15	150	yes	3.3%	0.971
ionosphere	15	150	no	12.0%	0.976
	15	150	yes	12.8%	0.978
lung cancer	15	150	no	0.0%	0.999
	15	150	yes	0.0%	1.000
wine	15	150	no	8.8%	0.987
	15	150	yes	8.8%	0.992

occurrence, the columns in table 1 denote the dataset, the network size measured in hidden units, the committee size i.e. the number of weights from the posterior used for prediction and whether or not we have used ARD of inputs. The last two columns denote the generalization error and the mean of the margin distribution.

In order to illustrate the correlation between the margins and the generalization performance, we present the scatter plot shown in figure 3. This plot shows the mean of the margin distributions versus generalization performance. We see a strong positive correlation (correlation coefficient of 0.896) among generalization performance and margins.

We also tried to link generalization performance and margin distributions within one data set, but the result was disappointing. We

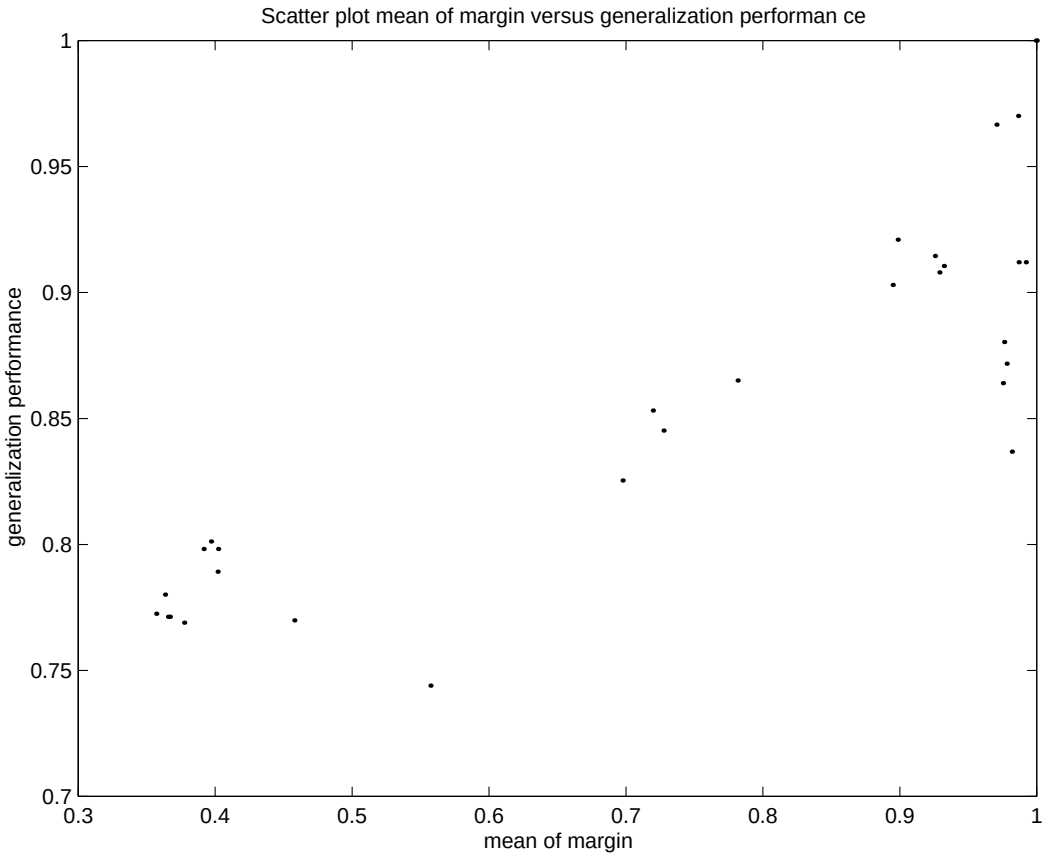

Figure 3: This scatter plot shows the generalization performance versus the corresponding mean of the margin distribution for all experiments. We see a strong correlation of 0.896, which is due to different difficulties of the data sets.

made twenty one pairwise comparisons of different algorithmic settings, which resulted in five significant different performance results. In one of these cases, the margin distribution showed no significant differences. In one case the margin distribution suggested the correct trend in generalization performance, but three of the comparisons showed the wrong trend. What are the reasons for the difficulties to correlate the margin distribution with generalization performance? Except for the two experiments with exactly the same number of hidden units that use different numbers of samples from the posterior, all experiments were performed with different settings of the classifiers or prior specifications. These changes lead to hypothesis classes with different complexities. Therefore the bounds between the margins on the training data and the margin we expect in the generalization case are different. Classifiers with lower complexity will in general show smaller margins, but at the same time the margins observed on new data will also shrink. This makes of course any conclusions from the margins observed on training data difficult.

5 Conclusions

Our theoretical analysis shows that Bayesian Classifiers of the kind described in [10] 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 translate into lower effective VC dimensions 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 over-complex systems.

Extensive experimental results on real world data have confirmed the theoretical predictions by exhibiting margin distributions which are concentrated around the maximal value.

The main aim of this paper has been to co-locate Bayesian Classifiers in the same category of two other learning systems – namely Support Vector Machines and Adaboost – which were motivated by very different considerations but which exhibit very similar behaviours with respect to overfitting. A unified analysis of the three systems is now possible, which can make potentially fruitful comparisons and cross-fertilisations much easier.

Acknowledgements

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

John Shawe-Taylor has received support from the ESPRIT Working Group in Neural and Computational Learning 2, (NeuroCOLT2 Nr. 27150).

P. Sykacek is funded by the European Commission (Biomed-2 project SIESTA, grant BMH4-CT97-2040). The Austrian Research Institute for Artificial Intelligence is funded by the Austrian Federal Ministry 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 program, organised at the institute between August and December 1997. This joint work was initiated there.

The authors would also like to thank Chris Williams for useful discussions and to Radford Neal for making his code freely available over the Internet.

References

[1] Peter L. Bartlett, "The Sample Complexity of Pattern Classification with Neural Networks: The Size of the Weights is More Important than the Size of the Network," *IEEE Trans. Inf. Theory*, **44**(2), 525–536, (1998).

[2] Peter Bartlett and John Shawe-Taylor, Generalization Performance of Support Vector Machines and Other Pattern Classifiers, *In* 'Advances in Kernel Methods - Support Vector Learning', Bernhard Schölkopf, Christopher J. C. Burges, and Alexander J. Smola (eds.), MIT Press, Cambridge, USA, 1998.

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

[4] C. Cortes and V. Vapnik, Support-Vector Networks, *Machine Learning*, 20(3):273-297, September 1995

[5] Nello Cristianini, John Shawe-Taylor, and Peter Sykacek, Bayesian Classifiers are Large Margin Hyperplanes in a Hilbert Space, In D.H. Fisher, Jr., editor, *Proceedings of International Conference on Machine Learning, ICML'97*, Nashville, Tennessee, July 1997. Morgan Kaufmann Publishers.

[6] Leonid Gurvits, A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In *Proceedings of Algo-*

rithm Learning Theory, ALT-97, and as NECI Technical Report, 1997.

[7] D. J. C. MacKay. The evidence framework applied to classification networks. *Neural Computation*, 4:720–736, 1992.

[8] D. J. C. MacKay. A practical Bayesian framework for backprop networks. *Neural Computation*, 4:448–472, 1992.

[9] David A. McAllester, Some PAC-Bayesian Theorems, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, 1998.

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

[11] Radford Neal, *Priors for Infinite Networks*, Technical Report CRG-TR-94-1, Dept. of Computer Science, University of Toronto, 1994.

[12] 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`

[13] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

[14] R. Schapire, Y. Freund, P. Bartlett, W. Sun Lee, Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods. In D.H. Fisher, Jr., editor, *Proceedings of International Conference on Machine Learning, ICML'97*, pages 322–330, Nashville, Tennessee, July 1997. Morgan Kaufmann Publishers.

[15] John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, Martin Anthony, Structural Risk Minimization over Data-Dependent Hierarchies, *IEEE Trans. on Inf. Theory*, **44**(5), 1926–1940, 1998.

[16] John Shawe-Taylor and Nello Cristianini, Margin Distribution Bounds on Generalization, (submitted to EuroColt 99). NeuroCOLT Technical Report NC-TR-1998-020, 1998.
(`http://www.neurocolt.com`).

[17] John Shawe-Taylor and Robert Williamson, A PAC Analysis of a Bayesian Estimator, *Proceedings of Tenth Annual Conference on Computational Learning Theory*, 1997, pp. 2–9.

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

[19] Vladimir N. Vapnik, Esther Levin and Yann Le Cunn, Measuring the VC-dimension of a learning machine, *Neural Computation,* 6:851–876, 1994.

[20] Chris Williams, *Computation with Infinite Networks*, In M. C. Mozer and M. I. Jordan and T. Petsche, editors, *Advances in Neural Information Processing Systems 9, NIPS 96*, Denver, CO, December, 1996, MIT Press, 1997.