1 Bayesian Voting Schemes as Large Margin Classifiers

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

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

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. 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.

1.1 Introduction

In recent years, a new method for training neural networks has been proposed and used, mainly due to the work of MacKay and Neal [7, 8, 9]. The systems inspired by this approach are generally know as Bayesian Learning Algorithms and have proven to be quite resistant to overfitting.

Voting Scheme

They are characterized by the fact that they output 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.

Occam paradox

The classifiers obtained with this method are known to be highly 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" [10, 16], meaning by this networks whose number of tunable parameters is much larger than the sample size.

Hyperplane

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).

Large margin

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 [13]. 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.*[12], whose thoerems we will quote for reference.

These results not only explain the remarkable 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.

Experimental results confirming the predictions of our model are reported in a companion paper [5].

1.2 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 to a set of parametrized functions $f(x, \lambda)$, possibly encoding some prior knowledge. In the following we will denote by f_λ the hypothesis associated to the function $f(x, \lambda)$.

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.

Bayes classifiers

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 $e^{-\mathrm{loss}(f_\lambda))}$, 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 [9]. The most important fact about Bayesian algorithms is that they turn out to be quite resistant to overfitting [11, 9], to the point that it is possible to use networks larger than the number of training example, and to combine them in large committees. They are interesting not only because they work, but also because their behaviour seems to challenge intuition.

1.3 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 understand the properties of their margin, and so of their effective VC dimension. This concept was introduced by Vapnik *et al.* [15], though we use the term to mean the fat shattering dimension measured at the scale of the observed margin. Theorem 1.1 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 $\{-1, +1\}$.

Convex hull of function space

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 1.1

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* $\mathrm{VCdim}(H)$ of the set H the size of the largest shattered set, if this is finite or infinity otherwise.

Definition 1.2

Let H be a set of real valued functions. We say that a set of points X is γ*-shattered by* 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 $f_b \in H$ 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* fat_H of the set 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 [13] which involves the fat shattering dimension of the space of functions.

Theorem 1.1

Fat VC bound

Consider a real valued function class $\mathcal{H}$ having fat shattering function bounded above by the function $\mathrm{afat} : \mathbb{R} \rightarrow \mathcal{N}$ which is continuous from the right. Fix $\theta \in \mathbb{R}$. Then with probability at least $1 - \delta$ a learner who correctly classifies ℓ

independently generated examples $\mathbf{z}$ with $h = T_\theta(f) \in T_\theta(\mathcal{H})$ such that $\mathrm{er}_{\mathbf{z}}(h) = 0$ 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 $k = \mathrm{afat}(\gamma/8)$.

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.* [13]. 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 this volume [2].

Theorem 1.2

[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

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

Linear functions in a Hilbert space

In order to apply Theorems 1.1 and 1.2 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

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

with the inner product

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

There is a natural embedding of the input space X onto the unit sphere of $\mathcal{H}$ given by $\mathbf{x} \mapsto (f(\mathbf{x},\cdot) \mapsto f(\mathbf{x},\lambda))$, 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 functionals considered are those determined by the posterior distribution. The norm is given by

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

Hence,

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

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 Boolean valued hypothesis space is $e^{-\mathrm{r}\sigma}$, 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.

1.3.1 The 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 1.3

Let s_λ 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 $B_\lambda = \ell - 2\mathrm{s}_\lambda$, where ℓ is the sample size. Hypotheses having the same value of 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 risk of f_λ.

During the next proof we will need to know the probability in the prior distribution of hypotheses in our parameter space which have 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 ℓ.

Neutral prior

We make the further assumption that the prior probability of hypotheses which have empirical risk $\epsilon = r/\ell$ 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 for functions realising different patterns of r errors is $2^{-\ell}$. We will assume that 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 1.3

Under the above assumptions 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,

$$< M > = \frac{1}{\ell} \sum_{j \in J} a_j P_j \sum_{i \in S} y_i f_j(\mathbf{x}_i), = \frac{1}{\ell} \sum_{j \in J} a_j P_j B_j \qquad (1.1)$$
$$= \sum_{j \in J} a_j P_j (1 - 2\epsilon_j) = 1 - 2 \sum_{j \in J} a_j P_j \epsilon_j,$$

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

$$\sum_{j \in J} a_j P_j \epsilon_j = \frac{1}{2^\ell} \sum_{r=0}^{\ell} A C^r \binom{\ell}{r} \frac{r}{\ell}. \qquad (1.2)$$

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 $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_\ell)$, we can realise the mapping $\mathbf{x}_i \mapsto \mathbf{z}_i, i = 1, \ldots, \ell$, 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 $\mathbf{z}$ 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 1.3 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.

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.* [12], though they appear in a more general form in [1].

Following [12], 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 1.4

VC bound

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 $\theta > 0$:

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

Theorem 1.5

Finite H bound

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 $\theta > 0$:

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

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 1.5 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 1.4 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 (1.1)). 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 original approach to show how this can take advantage of a beneficial prior.

1.3.2 The effect of the prior distribution on the margin bound

Non neutral 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 $\rho < 1$. We first repeat the calculations of Theorem 1.3 for this case. The sum (1.2) must take into account that in this case

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

The factor $(1+\rho)^\ell$ cancels and the factor ρ appears wherever C appears, that is

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

while

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

Hence, we have shown the following generalization of Theorem 1.3.

Theorem 1.6

Under the above assumptions of a beneficial prior the margin of the Bayes Classifier $F(x) \in \mathcal{C}(H)$ is given by

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

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 1.2 at the appropriate scale. This bound on the effective VC dimension becomes,

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

where to keep the formulae simple we have ignored the factor of 64 arising for the scale $\gamma/8$ in Theorem 1.1.

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 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 $\rho = 1$, that is the uninformative prior considered in Section 1.3.1. 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 $1 - 2/\sqrt{\ell}$, though of course the bound 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)

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

Effect of prior

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.

1.4 Conclusions

Our theoretical analysis shows that Bayesian Classifiers of the kind described in [9] 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 overcomplex systems.

Unified framework

The main theoretical 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 and cross-fertilizations much easier.

Experimental results confirming the predictions of the model on some benchmark problems can be found in [5].

Acknowledgements

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

John Shawe-Taylor was supported in part by the EPSRC research grant number GR/K70366.

We wish to thank Chris Williams and Peter Sykacek for many useful discussions about Bayesian Learning Algorithms.

References

1. Peter Bartlett, Pattern Classification in Neural Networks, IEEE Transactions on Information Theory, to appear.
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 Shavlik, J., ed., *Machine Learning: Proceedings of the Fifteenth International Conference*, Morgan Kaufmann Publishers, San Francisco, CA.
6. Leonid Gurvits, A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In *Proceedings of Algorithm 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. Radford Neal, *Bayesian Learning in Neural Networks*, Springer Verlag, 1996
10. Radford Neal, *Priors for Infinite Networks*, Technical Report CRG-TR-94-1, Dept. of Computer Science, University of Toronto, 1994.
11. 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`
12. 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.
13. John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, Martin Anthony, Structural Risk Minimization over Data-Dependent Hierarchies, to appear in *IEEE Trans. on Inf. Theory*, and NeuroCOLT Technical Report NC-TR-96-053, 1996.

(`ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports`).

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

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

16. 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.