Non-symmetric Support Vector Machines

Jianfeng Feng

Sussex University, Brighton BN1 9QH, UK
http://www.cogs.susx.ac.uk/users/jianfeng

Abstract. A novel approach to calculate the generalization error of the support vector machines and a new support vector machine–non-symmatic support vector machine–is proposed here. Our results are based upon the extreme value theory and both the mean and variance of the generalization error are exactly ontained.

1 Introduction

Multilayer perceptrons, radial-basis function networks and support vector machines are three approaches widely used in pattern recognition. In comparison with multilayer perceptrons and radial-basis function networks, the support vector machine optimizes its margin of separation and ensures the uniqueness of the final result. It seems support vector machines have become established as a powerful technique for solving a variety of classification, regression and density estimation tasks[2]. In practical applications, it is also recently reported that the SVM outperforms conventional learning algorithms[1]. How much does the SVM improve a machine's capability of generalization? There are a few authors [2, 9, 3] who have carried out a detailed analysis on the performance of the SVM. Nevertheless, the exact behaviour of the SVM on generalization remains elusive: all results obtained up to date are upper bounds of the mean of the generalization error (see next section for definition).

Here we propose a novel approach in terms of the extremal value theory [8, 4, 5] to exactly calculate the generalization error of a SVM. Although we confine ourselves to the case of one dimension, the conclusions obtained are illuminating. Firstly the mean and variance (or distribution) of the generalization error are exactly calculated. In the literature, as we mentioned above, only upper bounds of the mean are estimated. Secondly our approach enables us to go a step further to compare different learning algorithms. We assert that the support vector machine does improve the generalization error, both mean and variance, by a factor of a constant. Thirdly we then further propose a new version of the SVM, called non-symmetric support vector machine, which could, in some circumstances, further reduce the mean of the generalization. The basic idea of the non-symmetric support vector machine is that to employ not only the support vectors which are the only information used in the SVM, but also the mean of samples. In fact the advantage of non-symmetric support machine could be easily understood. The essence of the SVM is to rely only on the set of samples

J. Mira and A. Prieto (Eds.): IWANN 2001, LNCS 2084, pp. 418–426, 2001.
© Springer-Verlag Berlin Heidelberg 2001

which take extremal values, the so-called support vectors. From the statistics of the extremal values, we know that the disadvantage of such an approach is that the information contained in most samples (no extremal values) is simply lost and is bound to be less efficient than an algorithm taking into the lost information. We refer the reader to [6] for more details.

2 Models

The models we are going to consider are the support vector machine and the worst learning machine. For the former, the basic assumption is that the learning is only dependent on the support vectors, and a maximization of the separation margin is fulfilled. For the later,we assume that after learning, the machine is only able to correctly recognize learned samples (see Fig. 1). Very different from most approaches in the literature, where the learning machines with high dimensional inputs are considered, here we consider only one dimensional case due to the following reasons. Firstly in the one dimensional case, we are able to carry out a rigorous calculation of the mean and variance of the generalization error. Secondly in the one dimensional case, we could fully understand why and how the support vector machine outperforms the worst learning machine and gain insights onto how to further improve the generalization capability of a learning machine (see Discussion).

Let us first introduce the model here. Suppose that the target function is $x = 0$, i.e. the correct separation function (target hyperplane, is $\mathrm{sign}(x)$. After learning t examples from $A_1 = \{x(i) > 0, i = 1, \cdots, t\}$ and $A_2 = \{y(i) < 0, i = 1, \cdots, t\}$, a new coming signal $\xi(t+1)$ is sampled from $U(0,1)$, the uniform distribution over $[0,1]$, with probability 1/2 and $U(-1,0)$ with probability 1/2. The generalization error is defined by

$$\epsilon(t) = P(0 \le \xi(t+1) \le x_0 | \mathcal{F}_t) I_{\{x_0 > 0\}} + P(0 \ge \xi(t+1) \ge x_0 | \mathcal{F}_t) I_{\{x_0 < 0\}} \quad (1)$$

where $\mathcal{F}_t$ is the sigma-algebra generated by $x(i), y(i), i = 1, \cdots, t$ and I_A is the indication function of the set A, i.e. $I_A(x) = 1$ if $x \in A$ and 0 otherwise.

Denote

$$x(tt) = \min\{x(i), i = 1, \cdots t\} \qquad y(tt) = \max\{y(i), i = 1, \cdots, t\}$$

for the SVM the separation hyperplane is given by

$$x_0 = \frac{x(tt) + y(tt)}{2}$$

for the worst learning machine the separation hyperplane is given by

$$x_0 = x(tt)$$

In the literature the expectation of $\epsilon(t)$ is called the generalization error. Here since we are able to calculate not only the mean of $\epsilon(t)$, but also the variance etc., we prefer to call $\epsilon(t)$ the generalization error, which is a random variable.

3 Generalization Errors: Symmetric Cases

The basic idea is to apply the extremal value theory in statistics to estimating the generalization error. To this end, we first introduce a lemma here[1].

Lemma 1. *Suppose that $x(i) \sim U(0,1)$, the uniform distribution over $[0,1]$, is identically and independently distributed for $i = 1, \cdots, t$. When $t \to \infty$ we have*

$$P(x(tt) \geq \frac{x}{t}) = \exp(-x), \qquad x > 0 \tag{2}$$

$$\langle x^k(tt)\exp(-\alpha x(tt))\rangle = \frac{k!}{(\alpha+1)^{k+1}t^k}, \qquad \alpha > 0, k = 1, 2, \cdots, \tag{3}$$

where $\langle \cdot \rangle$ denotes the expectation. In other words, the distribution density of $x(tt)$ is $h(x) = t\exp(-tx)$.

Proof. From example 1.7.9 in[8] we know that $P(\eta(tt) \leq 1 - x/t) = \exp(-x)$ for $\eta(tt)$ representing the largest maximum of $x(i)$. Then Eq. (2) is a simple consequence of the symmetry between 1 and 0 of the uniform distribution. In terms of Eq. (2) we have

$$\langle x^k(tt))\exp(-\alpha x(tt))\rangle = \int x^k t\exp(-(\alpha+1)tx)dx = \frac{k!}{(\alpha+1)^{k+1}t^k}$$

Lemma 1 simply tells us the asymptotic distribution of $x(tt)$ when t is large enough. Let $\alpha = 0$ in Eq. (3), we see that $\langle x^k(tt)\rangle$ conveges to zero with a rate of $1/t^k$. For a given random sequence $x(i)$, we could calculate its exact distribution rather than its asymptotic distribution, which will provide us with further information on its behaviour with small samples[7].

From now on we assume that both $x(i)$ and $y(i)$ are uniformly distributed random variables and will report further work in [7]. The generalization error of the SVM defined in the previous section can now be rewritten as a function of extremal values

$$\begin{aligned}\epsilon(t) &= \frac{1}{2}P(0 \leq \xi(t+1) \leq \frac{x(tt)+y(tt)}{2}I_{\{x(tt)+y(tt)>0\}}|\mathcal{F}_t) \\ &\quad + \frac{1}{2}P(0 > \xi(t+1) \geq \frac{x(tt)+y(tt)}{2}I_{\{x(tt)+y(tt)<0\}}|\mathcal{F}_t) \\ &= \frac{1}{2}\frac{x(tt)+y(tt)}{2}I_{\{x(tt)+y(tt)>0\}} - \frac{1}{2}\frac{x(tt)+y(tt)}{2}I_{\{x(tt)+y(tt)<0\}}\end{aligned}$$

Here we have used the fact that $x(i)$ is uniformly distributed. Due to the symmetry between $x(i)$ and $y(i)$ we further conclude that

$$\langle \epsilon^k(t)\rangle = \langle [\frac{x(tt)}{2}I_{\{x(tt)+y(tt)>0\}} + \frac{y(tt)}{2}I_{\{x(tt)+y(tt)>0\}}]^k\rangle \qquad k = 1, 2, \cdots \tag{4}$$

[1] In the sequence, we take the convention that all terms of order $O(\exp(-t))$ in an equality are omitted

We first consider the mean of the generalization error by calculating the mean of each term in the equation above. The first term is

$$\begin{aligned}\langle \frac{x(tt)}{2} I_{\{x(tt)+y(tt)>0\}} \rangle &= \langle \frac{x(tt)}{2} \int_{-x(tt)}^{0} t\exp(ty)dy \rangle \\ &= \langle \frac{x(tt)}{2} (1-\exp(-tx(tt))) \rangle \\ &= [\frac{1}{2t} - \frac{1}{8t}] = \frac{3}{8t}\end{aligned}$$

The second term turns out to be

$$\begin{aligned}\langle \frac{y(tt)}{2} I_{\{x(tt)+y(tt)>0\}} \rangle &= \langle \frac{y(tt)}{2} \int_{-y(tt)}^{1} t\exp(-tx)dx \rangle \\ &= \langle \frac{y(tt)}{2} [\exp(ty(tt)) - \exp(-t)] \rangle \\ &= -[\frac{1}{8t} - \frac{1}{2t}\exp(-t)]\end{aligned}$$

Therefore we have the following conclusion.

Theorem 1. *The mean of the generalization error of the support vector machine is*

$$\langle \epsilon(t) \rangle = \frac{1}{4t} \tag{5}$$

Although the proof of Theorem 1 is almost straightforward, it is very interesting to see the implications of its conclusion. In the literature, different upper bounds for the mean of the generalization error of the support vector machine have been found (see for example [9]). However, it seems the result of Theorem 1 is the first rigorous, and exact value of the mean. It is generally believed that the generalization error of the support vector machine is improved, in comparison with other conventional learning rules. How much does it exactly improve? We answer it in the following theorem.

Theorem 2. *For the worst learning machine, the mean of the generalization error is given by*

$$\langle \epsilon(t) \rangle = \frac{1}{2t}$$

Proof. Now the generalization error is simply given by

$$\epsilon(t) = x(tt)/2$$

which, combining with Lemma 1, implies the conclusion of the theorem.

It is well known in the literature that the mean of the generalization error of a learning machine decays at a rate of $O(1/t)$, independent of the distribution of input samples. The mean of the generalization error of both the support vector machine and the worst learning machine is of order $1/t$, as we could expect. The illuminating fact here is that the support vector machine improves the mean of the generalization error by a factor of $1/2$, in comparison with the worst learning machine. We want to emphasize here that the conclusion in Theorem 2 is independent of distributions, i.e. *universally* for the worst learning machine its generalization error is $1/t$ (see Lemma 3 in [5]for a proof). Nevertheless, for the support vector machine, the conclusions in Theorem 1 are obtained in terms of the assumptions of the uniform distribution of input samples. For any given distribution, we could calculate, as we developed in Theorem 1, its mean generalization error. The key and most challenging question is that whether the obtained conclusion is univeral, i.e. independent of input distribution, or not. A detailed analysis is outside the scope of the present letter and we will report it in [7].

In the literature the generalization error of the support vector machine is expressed in terms of the separation margin. We could easily do it here as well. Denote $d = x(tt) - y(tt)$ as the separation margin.

Theorem 3. *The mean of the generalization error of the support vector machine is*

$$\langle \epsilon(t) \rangle = \frac{\langle d \rangle}{8} \tag{6}$$

Proof. Since $\langle x(tt) \rangle = 1/t$, the conclusion follows.

So far we have known that in terms of the mean of the generalization error, the support vector machine improves the performance. How is the variance of the generalization error of the support vector machine, in comparison with conventional learning rules? To the best of our knowledge, there is no report in the literature to successfully calculate the variance of the generalization error. Due to Lemma 1, we have the following conclusions.

Theorem 4. *For the worst learning machine its variance of the generalization error is*

$$var(\epsilon(t)) = \frac{1}{4t^2}$$

For the support vector machine we have

$$var(\epsilon(t)) = \frac{1}{16t^2}$$

Proof. We only need to calculate $\langle \epsilon^2(t) \rangle$ of the support vector machine. For Eq. (4) we have

$$\begin{aligned} \langle x^2(tt) I_{\{x(tt)+y(tt)>0\}} \rangle &= \langle x^2(tt) \int_{-x(tt)}^{0} yt \exp(ty) dy \rangle \\ &= \langle x^2(tt)[1 - \exp(-tx(tt))] \rangle \\ &= [\frac{2}{t^2} - \frac{1}{4t^2}] = \frac{7}{4t^2} \end{aligned}$$

and

$$\begin{aligned}\langle y^2(tt) I_{\{x(tt)+y(tt)>0\}}\rangle &= \langle y^2(tt) \int_{-y(tt)}^{1} t\exp(-tx)dx\rangle \\ &= \langle y^2(tt)(\exp(ty(tt)) - \exp(-t))\rangle \\ &= [\frac{1}{4t^2} - \frac{2}{t^2}\exp(-t)]\end{aligned}$$

Furthermore

$$\begin{aligned}\langle x(tt)y(tt) I_{\{x(tt)+y(tt)>0\}}\rangle &= \langle x(tt) \int_{-x(tt)}^{0} yt\exp(ty)dy\rangle \\ &= \langle x(tt)[x(tt)\exp(-tx(tt)) - \int_{-x(tt)}^{0} \exp(ty)dy]\rangle \\ &= \langle x(tt)[x(tt)\exp(-tx(tt)) - \frac{1}{t}(1-\exp(-tx(tt)))]\rangle \\ &= [\frac{1}{4t^2} - \frac{1}{t^2} + \frac{1}{4t^2}]\end{aligned}$$

which gives the desired results.

In words, the support vector machine also improves the standard deviation of the generalization error by a factor of 1/2, comparing with the worst leaning machine. As aformentioned it seems results on $var(\epsilon(t))$ have not been reported in the literature.

We could go further to estimate the distribution density of the generalization error. However, from the fact that the mean and the standard deviation of the generalization error are equal to each other, we could guess that the distribution density of the generalization error is negatively distributed with the parameter $\langle\epsilon(t)\rangle$, for both the worst learning machine and the support vector machine, a conclusion which is proved in [7].

In summary, under some assumptions on its input distributions (see [7] as well), we grasp a complete picture of the generalization behaviour of the one diminsional support vector machine.

4 Generalization Error: Non-symmetric Cases

In the previous sections we have considered the support vector machine with symmetric input distributions. Certainly we do not expect that $x(i)$ and $y(i)$ are identically distributed in problems arising from practical applications. In this section we assume that $yL \sim U(-L, 0)$ and the generalization error is then

$$\epsilon(t) = \frac{1}{2}\frac{x(tt)+Ly(tt)}{2} I_{\{x(tt)+Ly(tt)>0\}} - \frac{1}{2L}\frac{x(tt)+Ly(tt)}{2} I_{\{x(tt)+Ly(tt)<0\}} \tag{7}$$

where $L > 0$ is a constant. The first term in Eq. (7) is

$$\begin{aligned}\langle\frac{x(tt)}{2} I_{\{x(tt)+Ly(tt)>0\}}\rangle &= \langle\frac{x(tt)}{2}\int_{-x(tt)/L}^{0} t\exp(ty)dy\rangle \\ &= \langle\frac{x(tt)}{2}(1-\exp(-tx(tt)/L))\rangle \\ &= [\frac{1}{2t} - \frac{L^2}{2(L+1)^2 t}]\end{aligned}$$

and the second term

$$\begin{aligned}\langle\frac{Ly(tt)}{2}I_{\{x(tt)+Ly(tt)>0\}}\rangle &= \langle\frac{Ly(tt)}{2}\int_{-Ly(tt)}^{1} t\exp(-tx)dx\rangle\\ &= \langle\frac{Ly(tt)}{2}(\exp(tLy(tt))-\exp(-t))\rangle\\ &= -[\frac{L}{2(L+1)^2t}-\frac{L}{2t}\exp(-t)]\end{aligned}$$

Similarly for the third and the fourth term we have

$$\begin{aligned}\langle\frac{x(tt)}{2}I_{\{x(tt)+Ly(tt)<0\}}\rangle &= \langle\frac{x(tt)}{2}\int_{-1}^{-x(tt)/L} t\exp(ty)dy\rangle\\ &= \langle\frac{x(tt)}{2}(\exp(-tx(tt)/L)-\exp(-t))\rangle\\ &= \frac{L^2}{2(L+1)^2t}\end{aligned}$$

and

$$\begin{aligned}\langle\frac{Ly(tt)}{2}I_{\{x(tt)+Ly(tt)<0\}}\rangle &= \langle\frac{Ly(tt)}{2}\int_{0}^{-Ly(tt)} t\exp(-tx)dx\rangle\\ &= \langle\frac{Ly(tt)}{2}(1-\exp(tLy(tt)))\rangle\\ &= -[\frac{L}{2t}-\frac{L}{2(1+L)^2t}]\end{aligned}$$

Summing together we finally obtain

$$\langle\epsilon(t)\rangle = \frac{L}{4(1+L)t}+\frac{1}{4(1+L)t}=\frac{1}{4t}=\frac{\langle d\rangle}{4(1+L)} \tag{8}$$

where $d = x(tt) - Ly(tt)$.

Eq. (8) tells us that the mean of the generalization error is the same as the symmatic case and is proportional to $1/(1+L)$, where $1+L$ is conventionllly thought of as the gap between support vectors. It is somewhat surprising to note that the mean of the generalization error of the support vector machine is independent of the scaling of the input distribution. From the proof of Eq. (8) we see that the summation of the first and second term of Eq. (7) is not equal to the summation of the third and the fourth term, much as $\langle\epsilon(t)\rangle$ is independent of L. This reveals one of the difficulties to prove a general conclusion as developed in [5]. We have obtained a general conclusion and we will report it in [7].

5 Non-symmetric SVM

For the nonsymmetric case considered in the previous section, if the separation hyperplane is again $[x(tt)+y(tt)]/2$ then the generalization error is

$$\epsilon(t) = \frac{1}{2}\frac{x(tt)+y(tt)}{2}I_{\{x(tt)+y(tt)>0\}} - \frac{1}{2L}\frac{x(tt)+y(tt)}{2}I_{\{x(tt)+y(tt)<0\}}$$

From Theorem 1 we know that

$$\langle \epsilon(t) \rangle = \frac{1}{8t} + \frac{1}{8tL}$$

When L is large we then have

$$\langle \epsilon(t) \rangle = \frac{1}{8t}$$

a further reduction of the mean of the generalization error is achieved. The similar result is true for the variance of the generalization error.

The idea above can be implememted in the following way. We assume that A_1 and A_2 are the data set to be learnt.

1. According to the support vector machine algorithm, we obtain the separation hyperplane s_1.
2. Calculate the distance beween the mean of A_1, A_2 and s_1, denoting as d_1 and d_2 respectively. When L is large, we have $d_1 = 1/2 - [x(tt) + Ly(tt)]/2 \sim 1/2$ and $d_2 = L/2 + [x(tt) + Ly(tt)]/2 \sim L/2$.
3. In parallel with the hyperplane s_1, we find a new hyperplane s_2 so that

   $$\frac{c_1}{d_1} = \frac{c_2}{d_2} \tag{9}$$

   where c_1 and c_2 are the distance between s_2 and A_1 and A_2 respectively. We have $c_1 = x(tt) - [x(tt) + y(tt)]/2 = [x(tt) - y(tt)]/2$ and $c_2 = [x(tt) + y(tt)] - Ly(tt) \sim -Ly(tt)$. Hence when $s_2 = [x(tt) + y(tt)]/2$, we have Eq. (9).

Since in general the obtained separation hyperplane is not symmetric about the support vectors (maximization of the separation margin), we call s_2 the separation hyperplane of a non-symmetric support vector machine. The fact that the non-symmetric support vector machine improves the mean of the generalization error of s_1 could be easily understood. The support vector machine use only the information contained in the support vectors, while the non-symmetric support vector machine explore the information of the whole data set since the mean of the data is also taken into account.

6 Discussion

By virtue of the extremal value theory, we present here a novel approach to calculate the mean and variance of the generalization error of the support vector machine and the worst learning machine. The exact mean and variance of the generalization error are obtained. To estimate upper bounds for the support vector machine is currently a very active topic. Our results reveal, for the first time in ther literature, that how much the SVM improves the generalization error, comparing with other learning algorithms. Much as we consider a very simple case here, our results could also be used as a criteria to check how tight an estimated upper bound of general cases is (see [9, 3] and references therein). The extremal value theory is somewhat similar to the central limit theorem: a powerful and universal theorem and is almost independent of the sample distributions (see [8, 4, 5] for details). We hope the new techniques introduced here could help us to clarify some issues related to the SVM. Some of them we would like to further pursue in future publications are the following.

- The support vector machine improves the mean of the generalization error by a factor of a constant. From the calculations presented in the paper, we see that the next term in the mean of the generalization error is of order $\exp(-t)$. To find a learning algorithm with the mean of the generalization of order $\exp(-t)$–an exponential machine– would be a real breakthrough in the field. The approach presented here provides us with such a possibility.
- We only consider the case of one dimension. Certainly it is interesting to consider the models of high dimension. We will report it elsewhere[7].
- The extremal values are more sensitive to perturbations than other statistical quantities such as the mean or median of samples. It is therefore interesting to carry out a study on how the SVM relies on perturbations.

Acknowledgement. The work was partially supported by BBSRC and a grant of the Royal Society.

References

1. Brown M., Grundy W., Lin D., Cristianini N., Sugnet C., Furey T., Ares Jr. M., and, Haussler D.(1999) Knowledge-based Analysis of Microarray Gene Expression Data using Support Vector Machines. *Proceedings of the National Academy of Sciences* **97** 262-267.
2. Cristianini, N., and, Shawe-Taylor J. (2000) *An introduction to support vector machines* Cambridge University Press: Cambridge UK.
3. Dietrick R., Opper M., Sompolinsky H. (1999) Statistical mechanics of support vector networks *Phys. Rev. Letts*, **82** 2975-2978.
4. Feng, J.(1997) Behaviours of spike output jitter in the integrate-and-fire model. *Phys. Rev. Lett.* **79** 4505-4508.
5. Feng, J.(1998) Generalization errors of the simple perceptron. *J. Phys. A.* **31**, 4037-4048.
6. Feng, J., and Williams P. (2001) Calculation of the generalization error for various support vector machines *IEEE T. on Neural Networks* (in press).
7. Feng, J., and Williams P. (2001) Support vector machines-a theoretical and numerical study (in Prepartion)
8. Leadbetter, M.R., Lindgren, G. & Rootzén, H.(1983) *Extremes and Related Properties of Random Sequences and Processes,*Springer-Verlag, New York, Heidelberg, Berlin.
9. Vapnik V., and, Chapelle O.(2000) Bounds on error expectation for support vector machines *Neural computation* **12** 2013-2036.