Effects of Soft Margins on Learning Curves of Support Vector Machines

Kazushi IKEDA* Tsutomu AOISHI

Graduate School of Informatics, Kyoto University
Kyoto 606-8501 Japan

Abstract

The generalization properties of support vector machines (SVMs) are examined. From a geometrical point of view, the estimated parameter of an SVM is the one nearest the origin in the convex hull formed with given examples. Since introducing soft margins is equivalent to reducing the convex hull of the examples, an SVM with soft margins has a different learning curve from the original. In this paper we derive the asymptotic average generalization error of SVMs with soft margins in simple cases and quantitatively show that soft margins increase the generalization error.

Introduction

In recent years, support vector machines (SVMs) have attracted much attention as a new classification technique with good generalization ability in applications such as pattern classification [6, 15, 17–19]. The basic idea of SVMs consists of mapping input vectors into a high-dimensional feature space and separating the feature vectors linearly with the optimal hyperplane in terms of margins, i.e. distances of given examples from a separating hyperplane.

To assure convergence in linearly inseparable cases and to avoid overfitting to noisy data or outliers in examples, soft margins were introduced in SVMs which make them less sensitive to given examples by using slack variables for margin-constraint violation [6, 18, 19].

Theoretical background for the generalization ability of SVMs has been presented mainly in a framework of probably approximately correct (PAC) learning [6, 18, 19] where a kind of complexity of a learning-machine class called the VC dimension plays an important role [20]. Another criterion for measuring the generalization ability is the average generalization error [1–4, 10, 12] called a learning curve. Studies of the learning curves of kernel methods including SVMs are still being developed both from a statistical mechanical approach [7, 13, 14] and an asymptotic statistical approach [8, 9].

The former approach takes noise into account in terms of a finite temperature [13], not soft margins. The latter approach has, so far, never considered soft margins. Intuitively speaking, introducing soft margins increases the generalization error if the given problem is linearly separable although it is necessary in inseparable cases. In this paper, we quantify the effects of soft margins on the asymptotic generalization ability in simple cases, that is, the input space is one-dimensional.

Geometry of SVMs with Soft Margins

Here, we regard the feature vectors as the input vectors and consider a homogeneous linear dichotomy for brevity. That is, a separating function is represented by $\boldsymbol{w}'\boldsymbol{x}$ where $'$ denotes the transpose.

Suppose a set F_N of N examples $(\boldsymbol{x}_n, y_n)$, $n = 1, 2, \ldots, N$, is given. Then, since the margin of a separating hyperplane denoted by $\boldsymbol{w}$ is defined as the minimum distance between examples and the hyperplane, it is expressed as $\min_n \boldsymbol{w}'\boldsymbol{f}_n/\|\boldsymbol{w}\|$ where $\boldsymbol{f}_n = y_n\boldsymbol{x}_n$. The

*kazushi@i.kyoto-u.ac.jp

problem of finding $\hat{\boldsymbol{w}}$ that maximizes the margin is equivalent to the following optimization problem with linear inequalities,

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } \boldsymbol{w}'\boldsymbol{f}_n \geq 1, \quad n = 1, \ldots, N, \tag{1}$$

if the given examples are linearly separable. The dual problem is written as

$$\min_{\boldsymbol{\alpha}} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{n=1}^{N} \alpha_n\right]$$
$$\text{s.t. } \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \boldsymbol{f}_n, \quad 0 \leq \alpha_n, \tag{2}$$

using the Lagrangian multipliers $\alpha_n \geq 0$, $n = 1, \ldots, N$, where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)'$. To make it applicable to linearly inseparable data sets, slack variables ξ_n, $\xi = 1, \ldots, N$, have been introduced that allow the margin constraints to be violated as below:

$$\min_{\boldsymbol{w},\, \boldsymbol{\xi}} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n\right]$$
$$\text{s.t. } \boldsymbol{w}'\boldsymbol{f}_n \geq 1 - \xi_n, \quad \xi_n \geq 0, \tag{3}$$

where C is a given constant and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)'$.

Let us consider a rather general problem,

$$\min_{\boldsymbol{w}, \boldsymbol{\xi}, \beta} \left[\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n - \beta\right]$$
$$\text{s.t. } \boldsymbol{w}'\boldsymbol{f}_n \geq \beta - \xi_n, \quad \xi_n \geq 0. \tag{4}$$

This is equivalent to the ν-SVM proposed in [16]. It is obvious that this problem reduces to (3) if we fix β to unity and hence the solution of (3) is a suboptimal solution of (4). The dual problem of (4) is written as

$$\min_{\boldsymbol{w}} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } \boldsymbol{w} = \sum_{n=1}^{N} \alpha_n \boldsymbol{f}_n, \quad 0 \leq \alpha_n \leq C, \quad \sum_{n=1}^{N} \alpha_n = 1. \tag{5}$$

(5) means that the solution $\hat{\boldsymbol{w}}$ is the point nearest to the origin in the reduced convex hull of $\tilde{F}_N$ where $\tilde{F}_N$ is the set of the vectors $\boldsymbol{f}_n$, $n = 1, \ldots, N$.

The idea of reduced convex hulls has been introduced in [5]. Figure 1 shows examples of $\hat{\boldsymbol{w}}$ when $C = 1$ and $C < 1$, respectively. If $1/C$ is a positive integer M, the reduced convex hull of $\tilde{F}_N$ with $1/M$ is equivalent to the convex hull of ${}_N\mathrm{C}_M$ vertices each of which is a center of gravity of M distinct vectors in $\tilde{F}_N$ [11].

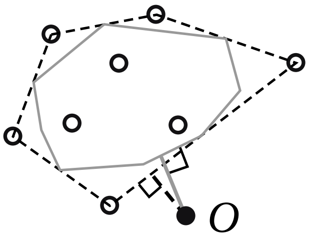

Figure 1: Points nearest to the origin O of the convex hull (the dashed line, $C = 1$) and the reduced convex hull (the gray solid line, $C = 1/2$) of examples shown by o's.

Average Generalization Error of SVMs

Let us show the learning curves of SVMs with soft margins in simple cases, provided that the input space is one-dimensional and that the given data is noise-free and linearly separable. The derivation is performed by using the fact that an SVM with soft margins given $\tilde{F}_N$ is equivalent to an SVM with hard margins given the centers of M vectors in $\tilde{F}_N$ as examples.

Fix the true parameter vector $\boldsymbol{w}^* = (0, \ldots, 0, 1)' \in S^m$ and assume that N inputs, $\boldsymbol{x}_n$, $n = 1, \ldots, N$, are independently uniformly chosen from S^m where S^m is an m-dimensional unit sphere. Then, the vectors $\boldsymbol{f}_n = y_n\boldsymbol{x}_n$, $n = 1, \ldots, N$, are uniformly distributed in S_+^m where $y_n = \mathrm{sgn}(\boldsymbol{w}^{*\prime}\boldsymbol{x}_n)$ and $S_+^m = \{\boldsymbol{f} | \boldsymbol{w}^{*\prime}\boldsymbol{f} \geq 0, \boldsymbol{f} \in S^m\}$. In this case, the probability that an estimate $\hat{\boldsymbol{w}}$ mispredicts the output of a new input $\boldsymbol{x}$ is written as θ/π where θ is the angle between $\hat{\boldsymbol{w}}$ and $\boldsymbol{w}^*$. In this paper, we define the average generalization error as the probability that an estimate $\hat{\boldsymbol{w}}$ mispredicts the output of a new input averaged over the given examples, which is often termed the prediction error. In the following subsections, we derive the average generalization error of SVMs for $m = 1$ in the asymptotic limit of $N \to \infty$.

Hard Margins' Case

When $M = 1$, the nearest point in the convex hull of examples is the midpoint of the two examples, denoted by $\boldsymbol{f}_{\mathrm{L}}$ and $\boldsymbol{f}_{\mathrm{R}}$, nearest to both endpoints of the semicircle S^1_+. That is, the SVM solution $\hat{\boldsymbol{w}}$ is written as $(\boldsymbol{f}_{\mathrm{L}} + \boldsymbol{f}_{\mathrm{R}})/2$ and its angle θ with $\boldsymbol{w}^*$ becomes $\theta = |\theta_{\mathrm{L}} - \theta_{\mathrm{R}}|/2$ where θ_{R} and θ_{L} are the angles of $\boldsymbol{f}_{\mathrm{L}}$ and $\boldsymbol{f}_{\mathrm{R}}$ with the endpoints.

Since the examples are independently chosen, the probability that the angle Θ of the nearest point with an endpoint is less than θ_{L} is written as

$$\mathrm{Prob}[\Theta \le \theta_{\mathrm{L}}] = 1 - (1 - \theta_{\mathrm{L}}/\pi)^N$$

and hence the density function is asymptotically

$$p(\theta_{\mathrm{L}}) \approx \frac{N}{\pi} \exp\left(-\frac{N}{\pi}\theta_{\mathrm{L}}\right)$$

where $\approx$ means that the ratio of both sides goes to unity when $N \to \infty$. Since θ_{R} has the same density function and the correlation between θ_{L} and θ_{R} can be neglected in the asymptotic limit of $N \to \infty$, the average generalization error $\epsilon_{\mathrm{g}}(N)$ is derived as

$$\epsilon_{\mathrm{g}}(N) \approx \frac{1}{2\pi} \int_0^A \int_0^A |\theta_{\mathrm{L}} - \theta_{\mathrm{R}}| p(\theta_{\mathrm{L}}) p(\theta_{\mathrm{R}}) \mathrm{d}\theta_{\mathrm{L}} \mathrm{d}\theta_{\mathrm{R}} \approx \frac{1}{2N}$$

where A is a certain constant which determines the domain of integration. Note that A does not affect the result in the asymptotic limit of $N \to \infty$.

Soft Margins' Case

In the case of a fixed $M \ge 2$, we consider the center of gravity of the nearest M examples from each endpoint. Let us denote by θ_l the angle between the $(l-1)$st and lth nearest examples to an endpoint. Then, the conditional probability density function of each angle $p(\theta_l|\theta_1, \ldots, \theta_{l-1})$ is asymptotically written as

$$\begin{aligned} & p(\theta_l|\theta_1, \ldots, \theta_{l-1}) \\ & \approx \frac{N-(l-1)}{\pi - \sum_{j<l}\theta_j} \exp\left(-\frac{N-(l-1)}{\pi - \sum_{j<l}\theta_j}\theta_l\right) \\ & \approx \frac{N}{\pi} \exp\left(-\frac{N}{\pi}\theta_l\right) \qquad (6) \end{aligned}$$

where the asymptotic equalities are based on the facts $\theta_j \ll \pi$ and $N \gg 1$. Let (6) be denoted by $p_l(\theta_l)$ since $p_l(\theta_l|\theta_1, \ldots, \theta_{l-1})$ does not depend on the conditions asymptotically. This property makes it possible to calculate the distribution functions of the angles θ_{R} and θ_{L} of $\boldsymbol{f}_{\mathrm{L}}$ and $\boldsymbol{f}_{\mathrm{R}}$ with the endpoints where $\boldsymbol{f}_{\mathrm{L}}$ and $\boldsymbol{f}_{\mathrm{R}}$ are the centers of M examples nearest to each endpoints. Since the SVM solution $\hat{\boldsymbol{w}}$ is written as $(\boldsymbol{f}_{\mathrm{L}} + \boldsymbol{f}_{\mathrm{R}})/2$ and its angle θ with $\boldsymbol{w}^*$ is $\theta = |\theta_{\mathrm{L}} - \theta_{\mathrm{R}}|/2$, we can derive the asymptotic average generalization error from the distributions of the angles θ_{R} and θ_{L}.

In the case of $M = 2$, for example, since the midpoint of the first and second nearest examples to an endpoint is a nearest center, the angle denoted by θ_{L} and its density $p(\theta_{\mathrm{L}})$ are written as

$$\begin{aligned} \theta_{\mathrm{L}} &= [\theta_1 + (\theta_1 + \theta_2)]/2, \\ p(\theta_{\mathrm{L}}) &\approx \frac{2N}{\pi} \exp\left(-\frac{N}{\pi}\theta_{\mathrm{L}}\right) - \frac{2N}{\pi} \exp\left(-\frac{2N}{\pi}\theta_{\mathrm{L}}\right), \end{aligned}$$

respectively. The density $p(\theta_{\mathrm{R}})$ of the other nearest center is calculated in the same way. Using these distributions, the asymptotic average generalization error is derived as

$$\epsilon_{\mathrm{g}}(N) = \left\langle \frac{|\theta_{\mathrm{L}} - \theta_{\mathrm{R}}|}{2\pi} \right\rangle = \frac{7}{12N}.$$

In the case of $M = 3$, the angle of a nearest center is written as $\theta_{\mathrm{L}} = (3\theta_1 + 2\theta_2 + \theta_3)/3$ and the asymptotic average generalization error is derived as

$$\epsilon_{\mathrm{g}}(N) = \left\langle \frac{|\theta_{\mathrm{L}} - \theta_{\mathrm{R}}|}{2\pi} \right\rangle = \frac{239}{360N}$$

in the same way.

Although the calculation is much more complicated, the asymptotic average generalization error for an arbitrary M can be derived using the density

$$p(\theta_{\mathrm{L}}) \approx \frac{N}{\pi(M-1)!}\sum_{i=1}^{M} i^{M-1} (-1)^{M-i}{}_{M}\mathrm{C}_i \exp\left(-\frac{MN}{\pi i}\theta_{\mathrm{L}}\right) \qquad (7)$$

as

$$\epsilon_{\mathrm{g}}(N) \approx \frac{1}{NM(M!)^2}\sum_{i,j=1}^{M} \frac{(-1)^{i+j} i^{M+2} j^{M}{}_{M}\mathrm{C}_{i}{}_{M}\mathrm{C}_j}{i+j}. \qquad (8)$$

For example, $\epsilon_{\mathrm{g}}(N) = 0.737/N$ for $M = 4$ and $\epsilon_{\mathrm{g}}(N) = 0.805/N$ for $M = 5$.

Asymptotic Analysis

Suppose $1 \ll M \ll N$, then an asymptotic analysis on M can be given as follows. The moment generating function $\phi(t)$ of $(\theta_{\mathrm{L}} - \theta_{\mathrm{R}})/\sqrt{M}$ is written as

$$\phi(t) = \left\langle \exp\left(t(\theta_{\mathrm{L}} - \theta_{\mathrm{R}})/\sqrt{M}\right)\right\rangle = \prod_{i=1}^{M} \frac{1}{1 - \dfrac{\pi^2 i^2}{N^2 M^3}t^2}. \qquad (9)$$

Hence the cumulant generating function $\psi(t) = \log\phi(t)$ is written as

$$\psi(t) = -\sum_{i=1}^{M}\log\left(1 - \frac{\pi^2 i^2}{N^2M^3}t^2\right) \qquad (10)$$

$$= t^2\sum_{i=1}^{M}\frac{\pi^2 i^2}{N^2M^3} + \mathrm{O}\left(M^{-1}\right) \qquad (11)$$

$$= \frac{\pi^2}{3N^2}t^2 + \mathrm{O}\left(M^{-1}\right). \qquad (12)$$

This means that $(\theta_{\mathrm{L}} - \theta_{\mathrm{R}})$ asymptotically obeys a normal distribution with mean 0 and variance $\dfrac{2\pi^2 M}{3N^2}$ when $M \to \infty$, therefore, the average generalization error for a large M is

$$\epsilon_{\mathrm{g}}(N) \approx \frac{\sqrt{M}}{\sqrt{3\pi}N} \qquad (13)$$

Conclusions and Discussions

The effects of soft margins on the generalization ability of SVMs have been examined. We derived the asymptotic average generalization errors in the simple noiseless case of $m = 1$ under the assumption that the parameter C, which represents the "softness" of margins, is the reciprocal of a positive integer M. The results show that soft margins increase the generalization errors. This can intuitively be interpreted as the fact that soft margins average given data and decrease the probability that points lie in the neighborhood of the separating hyperplane, whereas the probability that a test input is chosen from there remains constant. However, in general, averaging increases the signal-to-noise ratio and improves robustness against additive noise. Deriving the average generalization error in more general cases including noisy data will be the subject of future work.

References

[1] Amari, S., *Neural Networks*, **6** (1993), 161–166.

[2] Amari, S. et al., *Neural Computation*, **4** (1992), 605–618.

[3] Amari, S. and Murata, N., *Neural Computation*, **5** (1993), 140–153.

[4] Baum, E. B. and Haussler, D., *Neural Computation*, **1** (1989), 151–160.

[5] Bennett, K. P. and Bredensteiner, E. J., *Proc. ICML*, 2000, 57–64.

[6] Cristianini, N. and Shawe-Taylor, J., *An Introduction to Support Vector Machines*, Cambridge Univ. Press, 2000.

[7] Dietrich, R. et al., *Physical Review Letters*, **82** (1999), 2975–2978.

[8] Ikeda, K., *Proc.ICANN/ICONIP*, LNCS 2714, Springer, 2003, 201–208.

[9] Ikeda, K., *Trans. of IEICE*, **J86-D-II** (2003), 918–925.

[10] Ikeda, K. and Amari, S., *IEICE Trans. Fundamentals*, **E79-A** (1996), 409–414.

[11] Ikeda, K. and Aoishi, T., *Proc. FIT*, 2002, H-9.

[12] Opper, M. and Haussler, D., *Proc. COLT*, 1991, 75–87.

[13] Opper, M. and Urbanczik, R., *Physical Review Letters*, **86** (2001), 4410–4413.

[14] Risau-Gusman, S. and Gordon, M. B., *Physical Review E*, **62** (2000), 7092–7099.

[15] Schölkopf, B. et al., *Advances in Kernel Methods: Support Vector Learning*, Cambridge Univ. Press, 1998.

[16] Schölkopf, B. et al., *Neural Computation*, **12** (2000), 1207–1245.

[17] Smola, A. J. et al., *Advances in Large Margin Classifiers*, MIT Press, 2000.

[18] Vapnik, V. N., *The Nature of Statistical Learning Theory*, Springer, 1995.

[19] Vapnik, V. N., *Statistical Learning Theory*, John Wiley and Sons, 1998.

[20] Vapnik, V. N. and Chervonenkis, A. Y., *Theory of Probability and its Applications*, **16** (1971), 264–280.