ν-STRUCTURED SUPPORT VECTOR MACHINES

Sungwoong Kim, Jongmin Kim, Sungrack Yun, and Chang D. Yoo

Department of EE, Korea Advanced Institute of Science & Technology

sungwoong.kim01@gmail.com, {waterboy0309, yunsungrack}@kaist.ac.kr, cdyoo@ee.kaist.ac.kr

ABSTRACT

This paper considers a ν-structured support vector machine (ν-SSVM) which is a structured support vector machine (SSVM) incorporating an intuitive balance parameter ν. In the absence of the parameter ν, cumbersome validation would be required in choosing the balance parameter. We theoretically prove that the parameter ν asymptotically converges to both the empirical risk of margin errors and the empirical risk of support vectors. The stochastic subgradient descent is used to solve the optimization problem of the ν-SSVM in the primal domain, since it is simple, memory efficient, and fast to converge. We verify the properties of the ν-SSVM experimentally in the task of sequential labeling handwritten characters.

1. INTRODUCTION

Conventional C-style soft margin support vector machine (C-SVM) [1] introduces slack variables in order to relax the margin constraints for non-separable problems. The objective function of the C-SVM is the weighted sum of the regularization term and empirical margin risk term, which is the sum of slack variables, and a pre-defined constant C controls the trade-off between the margin maximization and empirical risk minimization. The constant C is an unintuitive parameter which is generally determined by time-consuming cross-validation.

The ν-support vector machine (ν-SVM) [2] substitutes C by an intuitive parameter ν which is an upper bound on the fraction of training margin errors and is also a lower bound on the fraction of support vectors. Therefore, it is possible to intuitively control the behavior of the trade-off between the margin maximization and empirical risk minimization. Steinwart [3] proved that a close upper estimate of twice the optimal Bayes risk is the asymptotically optimal choice of ν when an universal kernel is used. Recently, the ν-SVM has been further improved. Crisp et al. [4] suggested μ-SVM to penalize positive errors differently from negative based on the geometrical interpretations of the ν-SVM. They showed that the ν-SVM finds the hyperplane orthogonal to the closest line connecting the reduced convex hulls. Wu proposed the $\mu\nu$-SVM to control the misclassification cost ratio for an unbalance data [5] while Perez-Cruz et al. and Takeda st al. proposed the Eν-SVM to broaden the admissible interval of ν [6, 7].

In many real-world prediction problems such as natural language processing, computational biology, and computational vision, we often deal with complex structured outputs which involve trees, sequences, networks or sets. These structured outputs have inter-dependencies. Therefore, in contrast to the previous binary or multiway classifications which map each input instance into one of classes, the structured prediction is a generalized task to predict multiple structured labels simultaneously from multiple structured inputs. The structured support vector machine (SSVM) [8] is a large-margin learning framework based on a structured prediction model. The SSVM has been often shown to perform better than the conditional maximum likelihood estimator, since the SSVM increases the score margin calculated from the discriminant function by scaling it with a loss, and thus has better generalization ability for a structured prediction model.

In this paper, we consider the ν-SSVM which is formulated by replacing the balance parameter C in the SSVM with ν and incorporating ρ as the minimum margin. The parameter ν is proven here to be an upper bound on the empirical risk of margin errors and a lower bound on the empirical risk of the support vectors. Furthermore, both the lower and upper bound asymptotically converge to ν. Thus, under a large-margin learning framework for a structured prediction model, the ν-SSVM does not require cumbersome validation to determine the balance parameter as in SSVM. We solve the constrained optimization problem of the ν-SSVM using the stochastic subgradient descent algorithm which is simple, memory efficient, and fast to converge. The properties of the ν-SSVM are verified experimentally in the sequential labeling task of handwritten character.

The rest of the paper is organized as follows. Section 2 briefly reviews the ν-SVM and the SSVM. Section 3 describes the proposed ν-SSVM with important properties. Section 4 describes the stochastic subgradient descent algorithm which is adopted to the ν-SSVM. A number of experimental results are presented and discussed in section 5. Finally, Section 6 concludes this paper.

2. ν-SUPPORT VECTOR MACHINE AND STRUCTURED SUPPORT VECTOR MACHINE

In this section, we briefly review ν-SVM for binary classification and SSVM for structured prediction. The SSVM that we review incorporates not only margin scaling but also slack scaling. Only margin scaling formulation adopt the intuitive balance parameter ν.

2.1. ν-support vector machine

Let the decision function $h : \mathcal{X}(\subset \mathbb{R}^p) \rightarrow \mathcal{Y}(= \{\pm 1\})$ be the linear binary classifier, i.e.,

$$\hat{\mathbf{y}} = h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b), \tag{1}$$

where $\mathbf{x} \in \mathcal{X}$, $\hat{\mathbf{y}} \in \mathcal{Y}$, $\mathbf{w}$ and b are a p-dimensional input vector, predicted binary output, normal vector and bias, respectively.

This work was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Culture Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2010.

Given a set of m training samples, $\{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_m, \mathbf{y}_m)\}$, the conventional soft-margin SVM, called the C-SVM, relaxes the margin constraints by introducing slack variables ξ for non-separable problems:

$$\text{C-SVM:} \quad \min_{\mathbf{w},b,\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i \tag{2}$$
$$\text{s.t.} \quad y_i(\langle \mathbf{w}_i, \mathbf{x}_i\rangle + b) \geq 1 - \xi_i,$$
$$\xi_i \geq 0, \quad \forall i,$$

where $C > 0$ is the balance parameter which controls the trade-off between the model complexity and empirical risk. Since a pre-determined parameter C provides no intuition for selecting a proper value under the interval of $[0, \infty)$, C is generally determined by a cross-validation or validation on the development data. However, this is cumbersome and time-consuming.

The ν-SVM [2] is a variant of the soft-margin SVM, that incorporates an intuitively meaningful balance parameter by replacing C with ν and adding the minimum functional margin ρ to be maximized.

$$\nu\text{-SVM:} \quad \min_{\mathbf{w},b,\xi,\rho} \ \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \tag{3}$$
$$\text{s.t.} \quad y_i(\langle \mathbf{w}_i, \mathbf{x}_i\rangle + b) \geq \rho - \xi_i,$$
$$\rho \geq 0, \ \ \xi_i \geq 0, \quad \forall i,$$

where ν is the trade-off parameter as C in the C-SVM. Schölkopf et al. [2] showed that if the ν-SVM solution yields $\rho > 0$, the C-SVM with $C = 1/\rho$ produces the same solution. However, the ν-SVM has additional intuitive interpretations, e.g., ν is an upper bound on the fraction of margin errors and a lower bound on the fraction of support vectors. Thus, the ν-SVM formulation would be potentially more useful than the C-SVM formulation in real applications. Also, compared to $C \in [0, \infty)$, ν under the admissible interval $[\nu_{\min}, \nu_{\max}]$ where $0 \leq \nu_{\min} \leq \nu_{\max} \leq 1$ is easier to deal with [9]. In addition, the asymptotically optimal choice of ν under an universal kernel was proven to be a close upper estimate of twice the optimal Bayes risk [3]. Recently, the ν-SVM has been further improved by some variants, e.g., the μ-SVM [4], the $\mu\nu$-SVM [5], and the Eν-SVM [6, 7].

2.2. Support vector machine for structured output spaces

In the structured output space, a discriminant function $F : \mathcal{X}(\subset \mathbb{R}^p) \times \mathcal{Y}(\subset \mathbb{R}^q) \to \mathbb{R}$ is defined over input-output pairs considering dependencies between outputs. Given a p-dimensional input vector $\mathbf{x} \in \mathcal{X}$, a prediction is to find a q-dimensional output vector $\mathbf{y} \in \mathcal{Y}$ which maximizes F:

$$h(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathcal{Y}} F(\mathbf{x}, \mathbf{y}). \tag{4}$$

We assume that F is linear in the joint feature map $\Phi(\mathbf{x}, \mathbf{y})$ as

$$F(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y})\rangle, \tag{5}$$

where $\mathbf{w}$ is the parameter vector. The SSVM [8] optimizes $\mathbf{w}$ by minimizing a quadratic objective function subject to a set of linear soft margin constraints:

$$\text{SSVM:} \quad \min_{\mathbf{w},\xi} \ \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{m}\sum_{i=1}^{m}\xi_i \tag{6}$$
$$\text{s.t.} \quad \langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i, \mathbf{y})\rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i,$$
$$\xi_i \geq 0, \quad \forall i, \mathbf{y} \in \mathcal{Y}\backslash\mathbf{y}_i,$$

where $\Delta\Phi(\mathbf{x}_i, \mathbf{y}) = \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})$. In the SSVM, the margin is scaled with a loss $\Delta(\mathbf{y}_i, \mathbf{y})$, which is the measure of difference between a prediction $\mathbf{y}$ and a correct label $\mathbf{y}_i$ for ith sample. This means that the SSVM penalizes a margin violation in proportion to the loss associated to a competitive label by margin scaling.

3. ν-STRUCTURED SUPPORT VECTOR MACHINE

In order to acquire the same benefit attained in the ν-SVM by using an intuitive parameter ν, we modifies the SSVM and formulates the ν-SSVM as follows

$$\nu\text{-SSVM:} \quad \min_{\mathbf{w},\xi,\rho} \ \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \tag{7}$$
$$\text{s.t.} \quad \langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i, \mathbf{y})\rangle \geq \rho\Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i,$$
$$\xi_i \geq 0, \quad \forall i, \mathbf{y} \in \mathcal{Y}\backslash\mathbf{y}_i,$$
$$\rho \geq 0.$$

Note that ρ is incorporated into the margin constraint as the minimum margin and it is maximized in the objective function while training. Also, similar to the ν-SVM, ν substitutes C as a balance parameter between the margin maximization and empirical risk minimization. To derive the properties of the ν-SSVM, we transform the ν-SSVM to the following equivalent dual problem.

3.1. Dual problem

Applying the Lagrange multipliers $\alpha_{i\mathbf{y}} \geq 0$ for the margin constraint of ith sample and label $\mathbf{y}$, $\beta_i \geq 0$ for the ith slack variable, and $\delta \geq 0$ for ρ, the Lagrangian function can be formulated as

$$\begin{aligned} L(\mathbf{w}, \xi, \rho, \alpha, \beta, \delta) \\ = \ & \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{m}\sum_{i=1}^{m}\xi_i \\ & - \sum_{i,\mathbf{y}\neq\mathbf{y}_i} \alpha_{i\mathbf{y}}(\langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i, \mathbf{y})\rangle - \rho\Delta(\mathbf{y}_i, \mathbf{y}) + \xi_i) \\ & - \sum_{i}^{m}\beta_i\xi_i - \delta\rho. \end{aligned} \tag{8}$$

This function is minimized with respect to the primal variable $\mathbf{w}, \xi, \rho$ and simultaneously maximized with respect to the dual variables α, β, δ. To eliminate the primal variables, setting the corresponding partial derivatives to zero yields the following conditions:

$$\mathbf{w} = \sum_{i,\mathbf{y}\neq\mathbf{y}_i} \alpha_{i\mathbf{y}}\Delta\Phi(\mathbf{x}_i, \mathbf{y}), \tag{9}$$

$$\sum_{\mathbf{y}\neq\mathbf{y}_i} \alpha_{i\mathbf{y}} + \beta_i = \frac{1}{m}, \ \ \forall i, \tag{10}$$

$$\sum_{i,\mathbf{y}\neq\mathbf{y}_i} \alpha_{i\mathbf{y}}\Delta(\mathbf{y}_i, \mathbf{y}) - \delta = \nu. \tag{11}$$

Then, we obtain the following dual problem by substituting (9)-(11) into L:

$$\max_{\alpha} -\frac{1}{2}\sum_{i,\mathbf{y}\neq\mathbf{y}_i}\sum_{j,\bar{\mathbf{y}}\neq\mathbf{y}_j}\alpha_{i\mathbf{y}}\alpha_{j\bar{\mathbf{y}}}\langle\Delta\Phi(\mathbf{x}_i,\mathbf{y}),\Delta\Phi(\mathbf{x}_j,\bar{\mathbf{y}})\rangle$$
$$\text{s.t.}\quad 0\leq\sum_{\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\leq\frac{1}{m},\quad\forall i,$$
$$\sum_{i,\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}\Delta(\mathbf{y}_i,\mathbf{y})\geq\nu. \tag{12}$$

Similar to the dual problem of the ν-SVM in [2], in the dual problem of the ν-SSVM, there is an additional constraint including ν and the linear term $\sum_{i,\mathbf{y}\neq\mathbf{y}_i}\alpha_{i\mathbf{y}}$ does not appear in the objective function.

3.2. Properties of ν-structured SVM

The optimal slack variables can be written as

$$\xi_i=\left(\max_{\mathbf{y}\in\mathcal{Y}\setminus\mathbf{y}_i}\Big(-\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y})\rangle+\rho\Delta(\mathbf{y}_i,\mathbf{y})\Big)\right)_+,\forall i. \tag{13}$$

where $()_+$ represents a hinge loss. Since the optimal variables satisfy the following Karush-Kuhn-Tucker (KKT) condition,

$$\alpha_{i\mathbf{y}}(\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y})\rangle-\rho\Delta(\mathbf{y}_i,\mathbf{y})+\xi_i)=0,\forall i,\mathbf{y}\neq\mathbf{y}_i, \tag{14}$$

all optimal Lagrange multipliers, $\alpha_{i\mathbf{y}}$, for the margin constraints are zero except those related to the most competitive label for each sample:

$$\alpha_{i\mathbf{y}}=0,\quad\forall i,\mathbf{y}\neq\mathbf{y}_i^* \tag{15}$$

where the most competitive label $\mathbf{y}_i^*$ for ith sample is defined as

$$\mathbf{y}_i^*=\arg\max_{\mathbf{y}\in\mathcal{Y}\setminus\mathbf{y}_i}\Big(\langle\mathbf{w},\Phi(\mathbf{x}_i,\mathbf{y})\rangle+\rho\Delta(\mathbf{y}_i,\mathbf{y})\Big). \tag{16}$$

Therefore, the conditions (10) and (11) can be rewritten as

$$\alpha_{i\mathbf{y}_i^*}+\beta_i=\frac{1}{m},\ \forall i, \tag{17}$$
$$\sum_i\alpha_{i\mathbf{y}_i^*}\Delta(\mathbf{y}_i,\mathbf{y}_i^*)-\delta=\nu. \tag{18}$$

With another KKT condition, $\beta_i\xi_i=0,\forall i$, a training set can be divided into the following three subsets $\mathcal{M}_{(-,0,+)}$ according to the difference between the score of the correct label and the most competitive label:

- $\mathcal{M}_-=\{i:\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y}_i^*)\rangle<\rho\Delta(\mathbf{y}_i,\mathbf{y}_i^*)\}$ satisfying

$$\alpha_{i\mathbf{y}_i^*}=\frac{1}{m},\ \beta_i=0, \tag{19}$$
$$\xi_i=-\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y}_i^*)\rangle+\rho\Delta(\mathbf{y}_i,\mathbf{y}_i^*)>0, \tag{20}$$

- $\mathcal{M}_0=\{i:\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y}_i^*)\rangle=\rho\Delta(\mathbf{y}_i,\mathbf{y}_i^*)\}$ satisfying

$$0<\alpha_{i\mathbf{y}_i^*}<\frac{1}{m},\ \beta_i>0, \tag{21}$$
$$\xi_i=-\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y}_i^*)\rangle+\rho\Delta(\mathbf{y}_i,\mathbf{y}_i^*)=0, \tag{22}$$

- $\mathcal{M}_+=\{i:\langle\mathbf{w},\Delta\Phi(\mathbf{x}_i,\mathbf{y}_i^*)\rangle>\rho\Delta(\mathbf{y}_i,\mathbf{y}_i^*)\}$ satisfying

$$\alpha_{i\mathbf{y}_i^*}=0,\ \beta_i=\frac{1}{m},\ \xi_i=0, \tag{23}$$

where $\mathcal{M}_-$ is the set of samples which are error or lie within the margin and have positive slack variables. We denote samples in $\mathcal{M}_-$ by margin errors as in [2]. $\mathcal{M}_0$ is the set of samples which lie on the margin and $\mathcal{M}_+$ is the set of samples which lie over the margin. Similar to the properties of the ν-SVM [2], we can derive the following intuitive meaning of ν in the ν-SSVM.

Proposition 1 If $\rho>0$, then ν is an upper bound on the empirical risk of margin errors and a lower bound on the empirical risk of support vectors, .i.e.,

$$\frac{1}{m}\sum_{i\in\mathcal{M}_-}\Delta(\mathbf{y}_i,\mathbf{y}_i^*)\leq\nu\leq\frac{1}{m}\sum_{i\in\{\mathcal{M}_-\cup\mathcal{M}_0\}}\Delta(\mathbf{y}_i,\mathbf{y}_i^*), \tag{24}$$

where $\{\mathcal{M}_-\cup\mathcal{M}_0\}$ is the set of support vectors which have the positive α values.

Proof. One of the KKT conditions is that $\delta\rho=0$. This implies $\delta=0$, if $\rho>0$. Therefore, from eq. (18),

$$\nu=\sum_i\alpha_{i\mathbf{y}_i^*}\Delta(\mathbf{y}_i,\mathbf{y}_i^*) \tag{25}$$
$$=\sum_{i\in\mathcal{M}_-}\frac{1}{m}\Delta(\mathbf{y}_i,\mathbf{y}_i^*)+\sum_{i\in\mathcal{M}_0}\alpha_{i\mathbf{y}_i^*}\Delta(\mathbf{y}_i,\mathbf{y}_i^*). \tag{26}$$

Proposition 2 Suppose that $\rho>0$, that the training data $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^m$ were iid from $P(\mathbf{x},\mathbf{y})$ such that none of $P(\mathbf{x},\mathbf{y})$ contains any discrete component, and that $\Phi(\mathbf{x},\mathbf{y})$ is analytic and non-constant. With probability 1, asymptotically, ν equals to both the empirical risk of margin errors and the empirical risk of support vectors.

The proof is similar to the proposition 5.(iii) in [2]. Let $G(\mathbf{x},\mathbf{y})=\langle\mathbf{w},\Delta\Phi(\mathbf{x},\mathbf{y}^*)\rangle-\rho\Delta(\mathbf{y},\mathbf{y}^*)$ where $\mathbf{w}=\sum_i\alpha_{i\mathbf{y}_i^*}\Delta\Phi(\mathbf{x}_i,\mathbf{y}_i^*)$. Following a similar line to the proof of proposition 12.(iii) in [2], we can show that $\forall t\in\mathbb{R},\sup_G\hat{P}_m(|G(\mathbf{x},\mathbf{y})+t|=0)$ converges to zero in probability, where $\hat{P}_m$ is the sample-based estimate of P. Hence, using $t=0$ shows that almost surely, the number of samples in $\mathcal{M}_0$ tends to zero.

4. STOCHASTIC SUBGRADIENT DESCENT

The main obstacle in solving the constrained optimization problem of (7) is how to handle the exponentially large number of margin constraints. The SSVM algorithms generally incorporate an algorithm to reduce the number of constraints. For example, Tsochantaridis et al. [8] proposed the cutting plane algorithm, otherwise known as the column generation algorithm, that accumulates the most violating constraints in each iteration. On the other hand, to make a single margin constraint for each training sample, Sha et al. [10] applied the soft-max to approximate the hard-max while Ratliff et al. [11] iteratively performed a decoding to find the most competitive label and then optimizing parameters with respect to the most competitive label by a subgradient descent. Since, it is simple, memory efficient, and fast to converge, we use the stochastic subgradient descent [11] to solve the ν-SSVM in the primal domain.

Algorithm 1 Stochastic subgradient descent algorithm for the ν-SSVM

Choose: $\mathbf{w}_0$, ρ_0, $\{\mu_{\mathbf{w},t}\}_{t=1}^{\infty}$, and $\{\mu_{\rho,t}\}_{t=1}^{\infty}$

$t = 1$

repeat

Select $i(t)$ randomly.

Decode the most competitive label:

$$\mathbf{y}^*_{i(t)} = \arg\max_{\mathbf{y}\in\mathcal{Y}} \Big\{ \langle \mathbf{w}_{t-1}, \Phi(\mathbf{x}_{i(t)}, \mathbf{y})\rangle + \rho\Delta(\mathbf{y}_{i(t)}, \mathbf{y}) \Big\}.$$

Calculate subgradients:

$g_{\mathbf{w}_{t-1},i(t)} = \mathbf{w}_{t-1} - \Delta\Phi(\mathbf{x}_{i(t)}, \mathbf{y}^*_{i(t)})$

$g_{\rho_{t-1},i(t)} = \Delta(\mathbf{y}_{i(t)}, \mathbf{y}^*_{i(t)}) - \nu$

Update parameters by subgradient descent:

$\mathbf{w}_t = \mathbf{w}_{t-1} - \mu_{\mathbf{w},t} g_{\mathbf{w}_{t-1},i(t)}$

$\rho_t = \rho_{t-1} - \mu_{\rho,t} g_{\rho_{t-1},i(t)}$

Project $\mathbf{w}_t$ and ρ_t on to any additional constraints,

e.g., $\rho_t = 0$ if $\rho_t < 0$.

$t = t + 1$.

until convergence

Optimization criterion of the ν-SSVM in the primal domain in (7) can be rewritten as

$$\min_{\mathbf{w},\rho} \frac{1}{m} \sum_{i=1}^{m} f_i(\mathbf{w}, \rho) \tag{27}$$

where, by letting $\Delta(\mathbf{y}_i, \mathbf{y}) > 0,\ \forall \mathbf{y} \in \mathcal{Y}\backslash\mathbf{y}_i$, and $\Delta(\mathbf{y}_i, \mathbf{y}_i) = 0$,

$$\begin{aligned} & f_i(\mathbf{w}, \rho) \qquad (28)\\ &= \frac{1}{2}\|\mathbf{w}\|^2 + \max_{\mathbf{y}} \big(- \langle \mathbf{w}, \Delta\Phi(\mathbf{x}_i, \mathbf{y})\rangle + \rho\Delta(\mathbf{y}_i, \mathbf{y})\big) - \nu\rho. \end{aligned}$$

Thus, subgradients of f_i with respect to $\mathbf{w}$ and ρ are obtained as

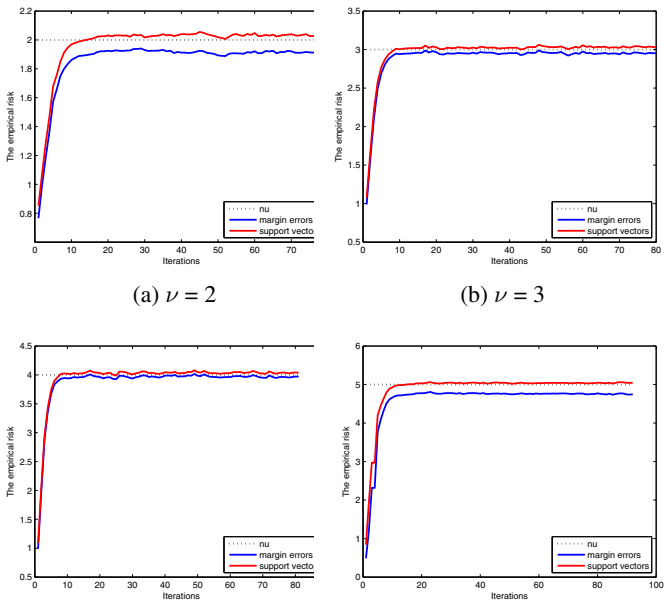
$$g_{\mathbf{w},i} = \mathbf{w} - \Delta\Phi(\mathbf{x}_i, \mathbf{y}^*_i), \tag{29}$$

$$g_{\rho,i} = \Delta(\mathbf{y}_i, \mathbf{y}^*_i) - \nu. \tag{30}$$

The subgradient descent algorithm for the ν-SSVM is summarized in Algorithm 1. In each iteration, we first decode the most competitive label $\mathbf{y}$ and calculate subgradients according to (29) and (30). The computational cost for inference can be greatly reduced if a decomposable joint map such as Markov structure and a decomposable loss such as Hamming distance are used. We then adjust parameters $\mathbf{w}$ and ρ by subgradient descent such that the large margin constraint is satisfied with respect to the most competitive label. Finally, updated $\mathbf{w}$ and ρ are projected on to any additional constraints if it is necessary. The theoretical convergence and robustness of the subgradient descent are analyzed [11], though in the ν-SSVM, $\mathbf{w}$ and ρ have to be jointly optimized.

5. EXPERIMENTS

To verify the properties of the ν-SSVM, handwritten character recognition, also called the optical character recognition, was performed. The data used for this experiment was provided by [12]. This data set contains 52152 characters, forming 6877 words. Each character is one of 26 English alphabets and has 128 binary pixel values. Principal component analysis was used to reduce the dimension from 128 to 20. For handwritten character recognition, a



(a) ν = 2 (b) ν = 3

(c) ν = 4 (d) ν = 5

Fig. 1. The evolutions of the empirical risk of margin errors and the empirical risk of support vectors according to different ν values.

first-order hidden Markov model was used as a structured model. The discriminant function based on this model was defined as a sum of transition scores and emission scores. The transition feature function for the transition score was an indicator function to represent the correlation between consecutive characters. The emission feature function was defined for each character label and based on the sufficient statistics of input to model a single unnormalized Gaussian distribution [10, 13]. The Hamming distance was used as a loss in the ν-SSVM, since the performance measure is a character error rate. As was done in [12], we divided whole data into 10 groups and conducted 10-fold cross validation.

In the experiments, our first goal is to verify Proposition 1 and 2 experimentally. Proposition 1 states that the parameter ν is an upper bound on the empirical risk of margin errors and a lower bound on the empirical risk of support vectors. We calculated the empirical risk of margin errors and the empirical risk of support vectors according to different ν values. Figure 1 shows that after a number iterations, the empirical risk of margin errors converges to below but near ν and simultaneously, the empirical risk of support vectors converges to above but near ν. Proposition 2 states that as the number of training data increases, the bounds of ν become tighter. Table 1 shows the change of both the empirical risk of margin errors and the empirical risk of support vectors as the number of training data increases, when ν = 2. Empirical risks were computed as averages of converged values after a number of iterations. As the number of training data increases, both bounds approach ν as in Proposition 2.

We additionally compared performances of the ν-SSVM to those of the SSVM when both were optimized by stochastic subgradient descent algorithm. Table 2 and 3 show averages of per-character test error rates according to different C and ν values,

Table 1. The empirical risk of margin errors and the empirical risk of support vectors according to the ratio between the number of training data and the number of test data (ν = 2). As the number of training data increases, the gap between two bounds gets narrower, forming more tight bounds.

training data(%)	10	30	50	70	90
empirical risk of support vectors	2.017572	2.010924	2.006138	2.004122	2.003841
empirical risk of margin errors	1.921715	1.937438	1.957724	1.969744	1.978293
bound gap	0.095857	0.073486	0.058414	0.034378	0.025548

respectively. As we mentioned in Section 3, both the parameter C and ν act as balance parameters. The best performance of the ν-SSVM, which was obtained when ν was 5, is comparable to that of the SSVM, which was obtained when C was 0.01.

In the SSVM, since the pre-determined parameter C has no intuition for selecting a proper value under the interval of $[0, \infty]$, C is generally determined by an exhaustive cross-validation, as shown in Table 2. In contrast, the parameter ν in our ν-SSVM formulation asymptotically converges to average empirical risk, which is formulated as the average loss over margin errors and support vectors. This property allows us to intuitively decide the upper value of the search interval for ν. For example, in handwritten character recognition task where the hamming loss is used, we can assume that ν is at least no larger than the average word length without much loss of generality. Since the handwritten character recognition data set has average word length of 8 characters, and hamming loss is used, average empirical loss might be less than 8. Therefore, it is highly probable that we get the optimal value for ν under the interval of $[0, 8]$.

However, it may be a natural question that even after the approximate search range is determined, we still need to do a cross validation to find the optimal value for ν, like the SSVM case. For the ν-SVM, Steinwart [3] performed several classifiers and picked the lowest error rate $\hat{R}_p$ to set it as the upper bound on the Bayes risk R_p. Then, depending on the confidence of tightness of $\hat{R}_p$, the author conducted a coarse search or fine search centered at $2\hat{R}_p$ to find the optimal value for ν. In this way, the author proved that incorporating prior knowledge on the Bayes risk for an effective search of ν yielded both better performance and shorter training times than the C-SVM. Motivated by [3], we could similarly consider an efficient way of choosing optimal value for parameter ν. We first run several classifiers and calculate the empirical risks of support vectors when $\rho = \rho_0$. Then, we could search around the minimum value of it for choosing optimal ν.

6. CONCLUSIONS

In this paper, we consider the ν-SSVM which is derived from the SSVM by replacing the balance parameter C in the SSVM with ν and incorporating ρ as the minimum margin. The parameter ν is proven to have an intuitive meaning that ν is an upper bound on the empirical risk of margin errors and is also a lower bound on the empirical risk of the support vectors. Moreover, the bounds of ν become tighter as the number of training data increases. We use the stochastic subgradient descent algorithm to solve the optimization problem of the ν-SSVM, since it is easy to optimize and fast to converge. The properties of the ν-SSVM are verified experimentally in the sequential labeling task of handwritten characters. Furthermore, an efficient method for setting the optimal value for parameter ν by incorporating some prior knowledge on the minimum empirical risk is considered. In this way, our ν-SSVM training is faster and more close to optimum since it does not require an exhaustive cross-validation to find the optimal balance parameter value as the SSVM does.

Table 2. Average per-character test error rate of the SSVM using 10-fold cross validation

C	0.001	0.01	0.1	1.0	10.0
test error rate (%)	14.36	13.85	14.43	14.47	14.56

Table 3. Average per-character test error rate of the ν-SSVM using 10-fold cross validation

ν	2	3	4	5	6
test error rate (%)	14.30	14.16	14.04	13.96	14.03

Acknowledgment

The authors would like to thank Dr. Alex Smola for valuable discussions on the ν-SSVM.

7. REFERENCES

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[2] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, pp. 1207–1245, 2000.

[3] I. Steinwart, "On the optimal parameter choice for ν-support vector machines," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1274–1284, 2003.

[4] D. J. Crisp and C. J. C. Burges, "A geometric interpretation of ν-svm classifiers," in *Advances in Neural Information Processing Systems 19*, 2000.

[5] X. Wu and R. Srihari, "New ν-support vector machines and their sequential minimal optimization," in *ICML*, 2003.

[6] F. Perez-Cruz, J. Weston, D. Herrmann, and B. Schölkopf, "Extension of the ν-svm range for classification," *Advances in Learning Theory: Methods, Models and Applications*, vol. 190, pp. 179–196, 2003.

[7] A. Takeda and M. Sugiyama, "ν-support vector machine as conditional value-at-risk minimization," in *ICML*, 2008.

[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and independent output variables," *Journal of Machine Learning Research 6*, 2005.

[9] C-C. Chang and C-J. Lin, "Training ν-support vector classifiers: theory and algorithms," *Neural Computation*, vol. 13, no. 9, pp. 2119–2147, 2000.

[10] F. Sha and L. K. Saul, "Large margin hidden markov models for automatic speech recognition," in *Advances in Neural Information Processing Systems 19*, 2007.

[11] N. Ratliff, J. A. Bagnell, and M. Zinkevich, "(online) subgradient methods for structured prediction," in *AISTATS*, 2007.

[12] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *Advances in Neural Information Processing Systems 16*, 2004.

[13] A. Gunawardana, M. Mahajan, A. Acero, and J .C. Platt, "Hidden conditional random fields for phone classification," in *Interspeech*, 2005.