CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 6**

Non-separable (soft) SVMs

Lecturer: Peter Bartlett *Scribe: Joseph Austerweil*

Outline of lecture:

1. Another geometric interpretation of hard-margin SVMs
2. Standard soft-margin SVM (C-SVM)
3. ν-SVM (interpretable reparameterization of C-SVM)

1 Another Geometric Interpretation of Hard-Margin SVMs

Previously, we explored the following definition of the SVM and the resulting geometric interpretation of its dual function (also shown in figure 1):

$$\begin{aligned} \min_{w \in \mathbb{R}^d} \quad & \|w\|^2 \\ \text{s.t.} \quad & \forall_i, y_i w' x_i \geq 1 \end{aligned}$$

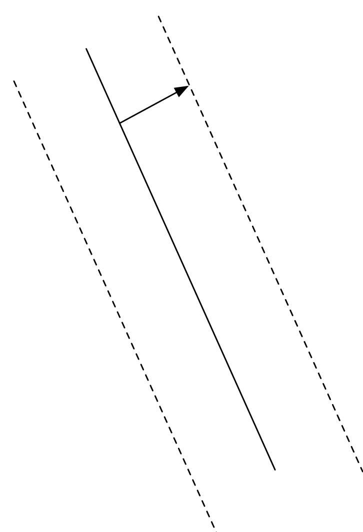

Figure 1: Hard SVM with its margins and decision boundary. The size of the margin is $\frac{1}{\|w\|^2}$

Instead, we can directly represent the margin, γ, which yields an equivalent optimization, but a different dual function.

$$\begin{array}{lll} \max_{w \in \mathbb{R}^d} & \gamma & \\ \text{s.t.} & \forall_i, y_i w' x_i \geq \gamma & \text{(dual parameter } \lambda_i) \\ & \|w\|^2 \leq 1 & \text{(dual parameter } \beta) \end{array}$$

We form the Lagrangian (switching the criterion to be a minimization of $-\gamma$):

$$L(w, \gamma, \lambda, \beta) = -\gamma + \sum_{i=1}^{n} \lambda_i (\gamma - y_i w' x_i) + \beta(\|w\|^2 - 1)$$

and at the minimum over w and γ, we have

$$\begin{aligned} \sum_i \lambda_i &= 1 \\ w &= \frac{1}{2\beta} \sum_i \lambda_i y_i x_i \end{aligned}$$

This gives the dual function,

$$g(\lambda, \beta) = -\frac{1}{4\beta} \sum_{i,j} \lambda_i \lambda_j y_i y_j x_i' x_j - \beta$$

and the dual optimization problem is

$$\begin{array}{ll} \min & \frac{1}{4\beta} \left\|\sum \lambda_i y_i x_i\right\|^2 + \beta \\ \text{s.t.,} & \sum \lambda_i = 1, \\ & \lambda_i \geq 0, \\ & \beta \geq 0 \end{array}$$

We can remove $\beta : \beta^2 = \frac{1}{4} \left\|\sum \lambda_i y_i x_i\right\|^2$. This results in the dual optimization:

$$\min \left\|\sum \lambda_i y_i x_i\right\| \text{s.t. } \lambda_i \geq 0, \sum_i \lambda_i = 1$$

Slater's condition implies strong duality.
We have:

$$w* = \frac{1}{2\beta} \sum_i \lambda_i y_i x_i = \frac{\sum_i \lambda_i y_i x_i}{\|\sum_i \lambda_i y_i x_i\|}$$

Or in other words, w* is the unit vector in the direction of the smallest norm element of the set

$$\text{co}(\{\sum_i y_i x_i : 1 \leq i \leq n\}) = \{\sum_i \lambda_i y_i x_i : \lambda_i \geq 0, \sum_i \lambda_i = 1\}.$$

From this formulation and Figure 2, we can observe that $w*$ points from the origin to the closest point on the convex hull formed by the positive and negative points (reflected through origin).

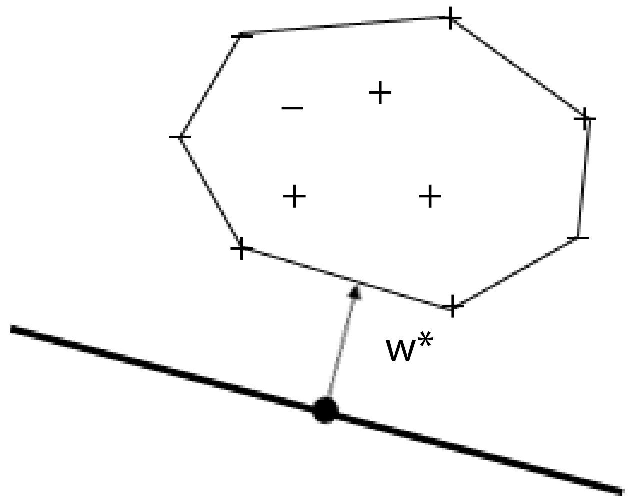

Figure 2: Optimal weight vector is from origin to closest point on convex hull formed from positive and reflected negative examples

2 Non-separable ("soft") SVMs

Non-separable SVMs allow the decision boundary to misclassify some examples, but it pays a cost for the number of violated constraints. We could form the optimization to be:

$$\min_{w\in\mathbb{R}^d} \|w\|^2 + \frac{C}{n}|S^c| \quad \text{s.t.} \quad \forall i \in S : y_i w' x_i \geq 1$$

$$\min_{w\in\mathbb{R}^d} \|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} 1[y_i w' x_i < 1]$$

However, this yields a nasty combinatorial optimization problem, so instead we replace the indicator function with a convex function.

$$\min \|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \phi(y_i w' x_i)$$

One possible function that is used for the soft SVMs of today's lecture (C-SVM and ν-SVM) is the hinge loss (see Figure 3):

$$\phi(\alpha) = (1-\alpha)_+ = \begin{cases} 1-\alpha & 1-\alpha > 0 \\ 0 & \text{otherwise} \end{cases}$$

Using the hinge-loss function, we form the primal optimization of the soft SVM to be:

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} (1 - y_i w' x_i)_+$$

$$\min_{w\in\mathbb{R}^d} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \quad \text{s.t., } \forall i : \underbrace{\xi_i \geq 0}_{\lambda_i} \quad \forall i : \underbrace{1 - \xi_i \leq y_i w' x_i}_{\alpha_i}$$

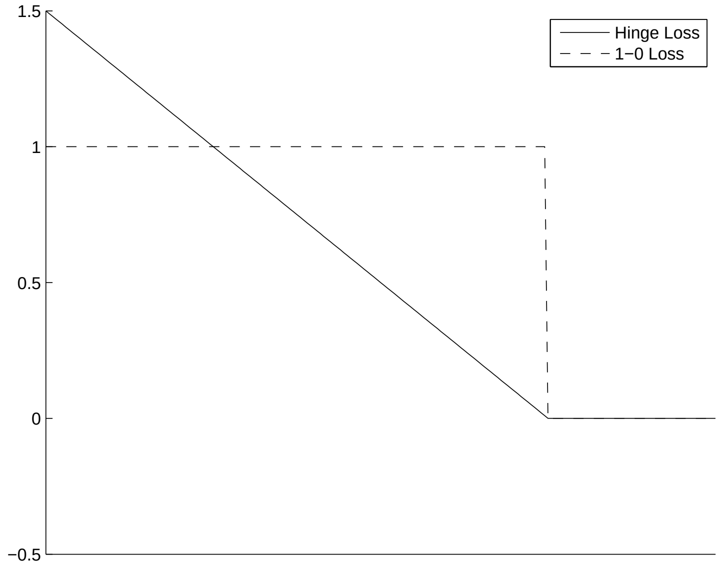

Figure 3: The hinge loss function (the solid line) is the typical loss function used for soft-margin SVMs.

C balances between the two parts of the criterion, so the larger the C the more we care about misclassified points. From this formulation, we can form the Lagrangian and derive the dual optimization:

$$L(w, \xi, \alpha, \lambda) = \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum \xi_i + \sum_i \alpha_i(1 - y_i w' x_i - \xi_i) - \sum_i \lambda_i \xi_i$$

Minimizing, we remove primal variables w and ξ from the optimization.

$$\frac{\partial L}{\partial w} = 0 \quad \Rightarrow \quad w = \sum_i \alpha_i y_i x_i$$
$$\frac{\partial L}{\partial \xi} = 0 \quad \Rightarrow \quad \alpha_i + \lambda_i = \frac{c}{n}$$

We form dual:

$$g(\alpha, \lambda) = \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j + \sum_i \alpha_i$$

Notice that we removed $\sum_i \xi(\frac{c}{n} - \alpha_i - \lambda_i)$ because $\forall i : \alpha_i + \lambda_i = \frac{c}{n}$, from the minimization. Thus, the form of the dual for the soft SVM is:

$$\begin{array}{cc} \max\limits_{\alpha,\lambda} & \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j \\ \text{s.t.} & \alpha_i \geq 0 \\ & \lambda_i \geq 0 \\ & \alpha_i + \lambda_i = \frac{C}{n} \end{array}$$

We can eliminate the λ_i variables, and replace the constraints with $0 \leq \alpha_i \leq \frac{C}{n}$. This constraint tells us that we cannot include too much weight on any point (at most $\frac{C}{n}$). In the hard margin case, we saw, via complementary slackness, that $\alpha_i > 0$ only when the corresponding example is on a margin. What is the similar condition for the soft-margin SVM?

- $\alpha_i > 0 \Rightarrow y_i w' x_i = 1 - \xi_i \leq 1$ (we are either at or on the wrong wide of the margin). The corresponding examples for $\alpha_i > 0$ are called the *support vectors*.

- $\underbrace{y_i x_i' w < 1}_{\text{"margin error"}} \Rightarrow \xi_i > 0$, and so $\lambda_i = 0 \rightarrow \alpha_i = \frac{C}{n}$

Note some examples that are classified correctly will still be considered a margin error and will have $\alpha = \frac{C}{n}$. Figure 4 shows this case. In the separable case, if C is greater than n times the largest α_i value, then the soft-margin SVM is equivalent to the hard-margin SVM.

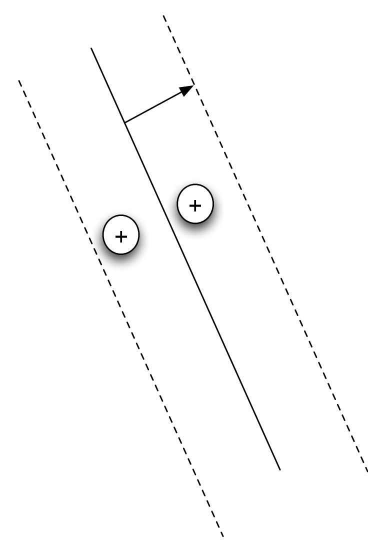

Figure 4: Both positive points, even though only one of which is misclassified, are considered margin errors and their corresponding α_i weight are $\frac{C}{n}$.

3 ν-SVM

The interpretation of C is not intuitive. We show that solving ν-SVM is an equivalent optimization problem, but ν has a more intuitive interpretation. We will show later that this can be understood as a reparamaterization of the C-SVM problem. We form ν-SVM:

$$\begin{array}{ll} \min\limits_{w,\rho} & \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n} (\rho - y_i w' x_i)_+ \\ \text{s.t.} & \rho \geq 0 \end{array}$$

Figure 5 shows the ν-SVM's decision boundary. An equivalent optimization problem (a quadratic program), stated in terms of slack variables, is

$$\begin{array}{cc} \min\limits_{w,\rho,\xi} & \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^n \xi_i \\ \text{s.t.} & \underbrace{\rho \geq 0}_{\gamma}, \\ & \underbrace{\xi_i \geq 0}_{\beta_i}, \\ & \underbrace{\xi_i \geq \rho - y_i w' x_i}_{\alpha_i} \end{array}$$

Using this definition, we can derive the Lagrangian and dual formulation.

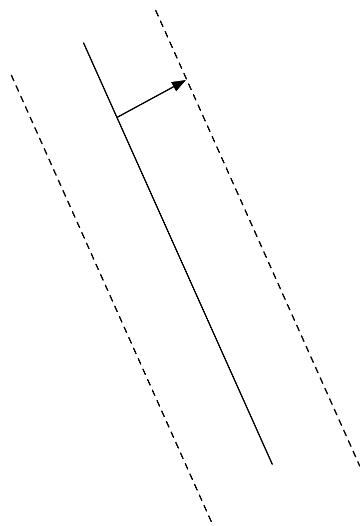

Figure 5: ν-SVM with its margins and decision boundary. The size of the margin is $\frac{\rho}{\|w\|}$

$$L(w,\rho,\xi,\alpha,\beta,\gamma) = \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_i \xi_i - \gamma\rho - \sum_i \xi_i\beta_i - \sum_i \alpha_i(y_i w' x_i + \xi_i - \rho)$$

Taking the minimum over our primal variables, w, ρ, and ξ, yields:

$$w = \sum_i \alpha_i y_i x_i \quad \nu = \textstyle\sum_i \alpha_i - \gamma \quad \beta_i + \alpha_i = \frac{1}{n}$$

This gives us the dual formulation:

$$\begin{array}{cc} \max\limits_{\alpha} & -\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i' x_j \\ \text{s.t.} & 0 \leq \alpha_i \leq \frac{1}{n}, \\ & \sum_i \alpha_i \geq \nu. \end{array}$$

Using the dual formulation, we can analyze the complementary slackness for the ν-SVM:

- $\alpha_i > 0 \Rightarrow y_i w' x_i = \rho - \xi_i \leq \rho$ (The corresponding vectors for $\alpha_i > 0$ are again called support vectors)

- $y_i w' x_i < \rho \Rightarrow \xi_i > 0 \Rightarrow \beta_i = 0 \Rightarrow \alpha_i = \frac{1}{n}$

Theorem 3.1. If $\rho > 0$ at solution, then:

$$\underbrace{|\{i : y_i w' x_i < \rho\}|}_{\#\text{ of margin errors}} \overset{(a)}{\leq} |\{i : \alpha_i = \frac{1}{n}\}| \overset{(b)}{\leq} \nu n \overset{(c)}{\leq} \underbrace{|\{i : \alpha_i > 0\}|}_{\#\text{ of support vectors}} \overset{(d)}{\leq} |\{i : y_i w' x_i \leq \rho\}|$$

Proof. (a) and (d) are given by complementary slackness.

$$\text{(b) } \rho > 0 \Rightarrow \gamma = 0 \Rightarrow \nu = \sum_i \alpha_i \geq \sum_i \alpha_i 1[\alpha_i = \frac{1}{n}] = \frac{1}{n} \sum_i 1[\alpha_i = \frac{1}{n}]$$

$$\text{(c) } \nu \leq \sum_i \alpha_i \leq \frac{1}{n} \sum_i 1[\alpha_i > 0]$$

□

By Theorem 3.1, we can think of νn as roughly the proportion of support vectors. Figure 6 shows the difference between the number of margin errors and the number of support vectors.

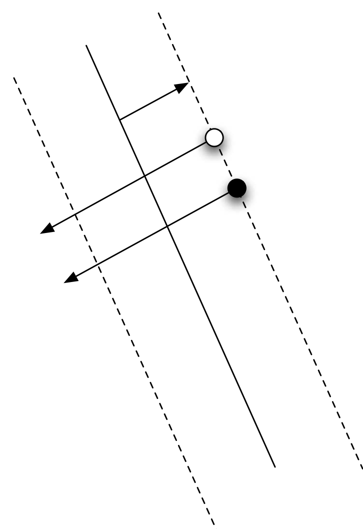

Figure 6: The decision boundary and margins once again for the ν-SVM. The open circle and arrow represent the margin errors whereas the closed circle represents the support vectors.

Theorem 3.2. If ν-SVM has a solution with $\rho > 0$, then C-SVM with $C = \frac{1}{\rho}$ gives an equivalent classifier.

Proof. If (w_*, ρ_*) is the solution to ν-SVM, we can fix $\rho = \rho_*$ and optimizing over w will not lead to a better value. That is, w^* is a solution to the optimization problem

$$\begin{aligned} \min_w \quad & \frac{1}{2}\|w\|^2 + \frac{1}{n}\sum_i \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0, \\ & y_i w' x_i \geq \rho^* - \xi_i. \end{aligned}$$

We can scale the objective by $1/{\rho^*}^2$ and the constraints by $1/\rho^*$ to obtain an equivalent optimization problem:

$$\begin{aligned} \min_{w} \quad & \frac{1}{2}\left\|\frac{w}{\rho^*}\right\|^2 + \frac{1}{n\rho^*}\sum_i \frac{\xi_i}{\rho^*} \\ \text{s.t.} \quad & \frac{\xi_i}{\rho^*} \geq 0, \\ & y_i \frac{w'}{\rho^*} x_i \geq 1 - \frac{\xi_i}{\rho^*}. \end{aligned}$$

And if we replace w/ρ^* with w and ξ_i/ρ^* with ξ_i, this is equivalent to the C-SVM with $C = \frac{1}{\rho_*}$. $\square$