CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 6**

Support Vector Machines for Non-separable Problems

Lecturer: Peter Bartlett *Scribe: Chuohao Yeo*

1 Soft-margin SVM

Previously, we have looked at using soft-margin SVM for problems in which the data is linearly non-separable. We choose the following optimization problem:

$$\min_w \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}(1 - y_i w^T x_i)_+ \tag{1}$$

where $(\alpha)_+ = \max(0, \alpha)$. C is a parameter that reflects how much we care about margin errors. We can think of the hard-margin SVM as using $C = \infty$.

We reformulated the above as a quadratic programming (QP) problem by introducing slack variables ξ_i:

$$\begin{aligned}
&\min_{w,\xi} \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\
&\text{s.t. } \xi_i \geq 0 \\
&\qquad y_i w^T x_i \geq 1 - \xi_i
\end{aligned}$$

Forming the dual problem, we get:

$$\begin{aligned}
&\max_\alpha \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
&\text{s.t. } 0 \leq \alpha_i \leq \frac{C}{n}
\end{aligned}$$

We can then write the primal solution in terms of the dual solution as $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$. Note that if $C = \infty$, we will get the same dual problem as in the hard-margin SVM.

1.1 Conditions from complementary slackness

We can obtain from complementary slackness the following conditions:

1. If $\alpha_i^* > 0$, then we must have that $1 - \xi_i^* - y_i w^{*T} x_i = 0 \Rightarrow y_i w^{*T} x_i \leq 1$. In other words, all support vectors are either on the wrong side of the decision boundary, or no more than $1/\|w^*\|$ away from the decision boundary.
2. If $\xi_i^* > 0$, (i.e. $y_i w_i^{*T} x_i < 1$), we must have $\lambda_i^* = 0$. Then, $\alpha_i^* = \frac{C}{n}$. In other words, if a vector x_i is a margin error (i.e. it is either on the wrong side of the decision boundary, or strictly less than $1/\|w^*\|$ away from the decision boundary), then $\lambda_i^* = 0$ and $\alpha_i^* = \frac{C}{n}$.

1.2 Choice of C

We can interpret a large C as penalizing margin errors more, with less concern on the size of $\|w\|^2$. If C is small, then the solution is less strongly influenced by a given data sample (x_i, y_i).

1.3 Geometric interpretation

In the case of hard-margin SVM in separable problems, we can view the optimization problem as finding:

$$\begin{aligned} w^* &= \arg\min_w \|w\|, w \in \text{convex hull}(y_i x_i) \\ &= \arg\min_w \|w\|, w \in \{\sum_{i=1}^n \lambda_i y_i x_i : 0 \leq \lambda_i \leq 1, \sum_{i=1}^n \lambda_i = 1\} \end{aligned}$$

For the case of soft-margin SVM, we have instead the following:

$$w^* = \arg\min_w \|w\|, w \in \{\sum_{i=1}^n \lambda_i y_i x_i : 0 \leq \lambda_i \leq D, \sum_{i=1}^n \lambda_i = 1\}$$

where D is some constant.

2 ν-SVM

The ν-SVM is an alternative formulation of the soft-margin SVM where the parameter ν lends itself more easily to interpretation than the parameter C from (1). We want to solve the following optimization problem:

$$\begin{aligned} &\min_{w,\rho} \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^n (\rho - y_i w^T x_i)_+ \\ &\text{s.t. } \rho \geq 0 \end{aligned}$$

Converting to a QP by introducing slack variables ξ_i, we obtain:

$$\begin{aligned} &\min_{w,\rho,\xi} \frac{1}{2}\|w\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^n \xi_i \\ &\text{s.t. } \rho \geq 0 \\ &\qquad \xi_i \geq 0 \\ &\qquad y_i w^T x_i \geq \rho - \xi_i \end{aligned}$$

The Lagrangian is obtained by introducing the Lagrange multipliers α, β and γ:

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

Minimizing the $L(\cdot)$ with respect to each of the primal variables, we have:

$$\nabla_w L(\cdot) = w - \sum_i \alpha_i y_i x_i = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i$$
$$\frac{\partial}{\partial \rho} L(\cdot) = -\nu - \gamma + \sum_i \alpha_i = 0 \Rightarrow \gamma = \sum_i \alpha_i - \nu$$
$$\frac{\partial}{\partial \xi_i} L(\cdot) = \frac{1}{n} - \beta_i - \alpha_i = 0 \Rightarrow \alpha_i + \beta_i = \frac{1}{n}$$

Substituting back into the Lagrangian, we have the dual function:

$$g(\alpha, \beta, \gamma) = -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

Hence, the dual problem is:

$$\begin{aligned} &\max_{\alpha} -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t. } 0 \le \alpha_i \le \frac{1}{n} \text{ ,(since } \beta_i \ge 0 \text{, and } \alpha_i + \beta_i = \frac{1}{n}) \\ &\qquad \sum_i \alpha_i \ge \nu \text{ ,(since } \gamma \ge 0 \text{, and } \gamma = \sum_i \alpha_i - \nu) \end{aligned}$$

2.1 Conditions from complementary slackness

1. If $\alpha_i^* > 0$, we must have that $y_i {w^*}^T x_i = \rho^* - \xi_i^* \le \rho^*$. In other words, all support vectors must be either on the wrong side of the decision boundary, or no more than $\rho^*/\|w^*\|$ away from it.
2. If $y_i {w^*}^T x_i < \rho^*$ (i.e. in the case of margin errors), then $\xi_i^* > 0$, which implies that $\beta_i^* = 0$, and $\alpha_i^* = \frac{1}{n}$.

2.2 Interpretation of ν

Theorem 2.1. If a solution has $\rho > 0$, then:

$$\begin{aligned} |\{i : y_i w^T x_i < \rho\}| &\le |\{i : \alpha_i = \frac{1}{n}\}| && (2) \\ &\le \nu n && (3) \\ &\le |\{i : \alpha_i > 0\}| && (4) \\ &\le |\{i : y_i w^T x_i \le \rho\}| && (5) \end{aligned}$$

Hence, from (3) and (5), we have that $|\{i : y_i w^T x_i < \rho\}| \le \nu n \le |\{i : y_i w^T x_i \le \rho\}|$. Thus, we can interpret ν as being approximately the fraction of data that are margin errors.

PROOF. (2) follows from the second complementary slackness condition. (3) comes from the fact that if $\rho > 0$, $\gamma = 0$ from complementary slackness, and so $\nu = \sum_i \alpha_i$. Thus, the number of vectors with $\alpha_i = \frac{1}{n}$ is at most νn. In addition, since $\alpha_i \le \frac{1}{n}$, we must have (4). Finally, (5) follows from the first complementary slackness condition. □

The following theorem tells us that ν-SVM is essentially the same as the soft-margin SVM.

Theorem 2.2. If there is a solution to the ν-SVM with $\rho > 0$, then the soft-margin SVM with $C = \frac{1}{\rho}$ gives the same decision rule.

PROOF. Suppose there is a solution to the ν-SVM with ρ^* and w^*. Then, w^* would solve the following optimization problem:

$$\begin{aligned} &\min_w \frac{1}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n \xi_i \\ &\text{s.t. } \xi_i \geq 0 \\ &\quad\quad y_i w^T x_i \geq \rho^* - \xi_i \end{aligned}$$

and hence also solves the following:

$$\begin{aligned} &\min_w \frac{1}{2}\|\frac{w}{\rho^*}\|^2 + \frac{1}{n\rho^*}\sum_{i=1}^n \frac{\xi_i}{\rho^*} \\ &\text{s.t. } \frac{\xi_i}{\rho^*} \geq 0 \\ &\quad\quad y_i(\frac{w^T}{\rho^*})x_i \geq 1 - \frac{\xi_i}{\rho^*} \end{aligned}$$

The above is just a scaled version of the soft-margin SVM problem, with $C = \frac{1}{\rho^*}$. □

3 Complexity regularization

We can rescale the objective function in (1) and write the optimization problem as:

$$\min_w \frac{1}{n}\sum_{i=1}^n l_{\text{hinge}}(y_i, w^T x_i) + \frac{\lambda}{2}\|w\|^2 \tag{6}$$

where $l_{\text{hinge}}(y_i, w^T x_i) = (1 - y_i w^T x_i)_+$. We can then interpret the first term as a measure of empirical risk, and the second term as a penalty on the complexity of the classifier. We can also view $l_{\text{hinge}}(\cdot)$ as a surrogate risk function.

3.1 Risk vs surrogate risk

Ideally, we would like to minimize the risk, $L(f) = \mathbb{E}l(y, f(x))$, where $l(y, f(x)) = 1[y \neq \text{sign}(f(x))]$. If the distribution is not known, we can also try to minimize the empirical risk: $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i))$. However, due to the combinatorial nature of this empirical risk minimization, it is a hard problem.

Instead, in soft-margin SVM, we consider minimizing an empirical ϕ-risk, $\hat{L}_\phi(f) = \frac{1}{n}\sum_{i=1}^n \phi(y_i, f(x_i))$, where $\phi(y, \hat{y}) = (1 - y\hat{y})_+$. In general, to formulate the minimization as a convex optimization problem, we would like $\phi(y, \cdot)$ to be convex (i.e. $\phi(y, \hat{u})$ should be convex in $\hat{y}$ with fixed y). For example, Adaboost uses $\phi(y, \hat{u}) = \exp(-Cy\hat{y})$, and neural networks use $\phi(y, \hat{y}) = (1 - y\hat{y})_+^2$.

We would like to answer the following question: does minimizing surrogate risk also minimize risk?

Definition. $L_\phi(f) = \mathbb{E}\phi(Y, f(X))$

We observe that $\phi(y, \hat{u}) \leq 1[y\hat{y} \leq 0] \Rightarrow L(f) \leq L_\phi(f)$.

Note that we can view the optimization problem in (6) as:

$$\begin{aligned} &\min_w \frac{1}{n}\sum_{i=1}^{n} \phi(y_i, w^T x_i) \\ &\text{s.t. } \|w\| \leq B \end{aligned}$$

for some suitably chosen B. Then, to show that minimizing the empirical surrogate risk is a reasonable thing to do, we will have to consider some uniform law of large numbers, with a w that is chosen uniformly from $\{w : \|w\| \leq B\}$. In other words, if

$$F = \{x \mapsto w^T x : \|w\| \leq B\}$$

we want to show that

$$\sup_{f \in F} |\hat{L}_\phi(f) - L_\phi(f)| \to 0$$

using some kind of uniform convergence of sample average to expectation.