CS281B/Stat241B. Statistical Learning Theory. Lecture 22.

Peter Bartlett

Overview

- Soft margin support vector machines
  - Quadratic program
  - Dual
  - ν-SVM

Recall: Hard Margin Support Vector Machine

$$\begin{aligned} \min_{\theta} \quad & \frac{1}{2}\|\theta\|^2 \\ \text{s.t.} \quad & y_i \theta^T x_i \geq 1, \qquad i = 1, 2, \ldots, n. \end{aligned}$$

$$f_n(x) = \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i^* y_i k(x_i, x)\right),$$

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2}\alpha^T \operatorname{diag}(y) K \operatorname{diag}(y) \alpha - \alpha^T 1 \\ \text{s.t.} \quad & \alpha \geq 0. \end{aligned}$$

Support Vector Machine

For the feasible region to be non-empty, there must be a θ with $y_i \theta^T x_i > 0$, i.e., all points classified correctly.

What if there is no such θ?

We could aim to minimize the proportion of constraints violated,

$$\frac{1}{n} \left| \left\{ i : y_i \theta^T x_i < 1 \right\} \right| ,$$

but this optimization problem is NP-hard.

Soft Margin Support Vector Machine

Instead, we can minimize a convex function of θ, such as

$$\min_{\theta} \qquad \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}(1 - y_i\theta' x_i)_+$$

where $(\alpha)_+ = \max\{\alpha, 0\}$.

Soft Margin Support Vector Machine

This is also a quadratic program:

$$\begin{aligned} \min_{\theta,\xi} \quad & \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \xi_i \geq 0 \\ & y_i \theta' x_i \geq 1 - \xi_i. \end{aligned}$$

Note: the optimal slack variables ξ_i satisfy

$$\xi_i = \left(1 - y_i \theta^T x_i\right)_+ .$$

Soft Margin Support Vector Machine

$$L(\theta, \xi, \alpha, \lambda) = \frac{1}{2}\|\theta\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - y_i\theta^T x_i - \xi_i) - \sum_{i=1}^{n}\lambda_i\xi_i$$

Minimizing over θ, ξ gives

$$\theta = \sum_i \alpha_i y_i x_i,$$

$$\frac{C}{n} = \alpha_i + \lambda_i,$$

so

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

Soft Margin Support Vector Machine

The dual problem is:

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

Eliminating the λ_i:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \alpha^T \operatorname{diag}(y) K \operatorname{diag}(y) \alpha - \alpha^T 1 \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq \frac{C}{n}. \end{aligned}$$

Support vectors

Note: the only change in going from the hard margin to the soft margin is the addition of the upper bound on the α_i.

Consider the consequences of complementary slackness:

$$\alpha_i^* \left(1 - y_i x_i^T \theta^* - \xi_i^*\right) = 0.$$

$$\lambda_i^* \xi_i^* = 0.$$

1. $\alpha_i^* > 0$ implies $y_i x_i^T \theta^* = 1 - \xi_i^* \leq 1$. That is, the 'support vectors' ($y_i x_i$ with $\alpha_i > 0$) are in the wrong halfspace $\{x : x^T \theta^* \leq 1\}$.
2. If $y_i x_i^T \theta^* < 1$, $\xi_i^* > 0$, so $\lambda_i^* = 0$, and $\alpha_i^* = C/n$. That is, the support vectors in the open halfspace $\{x : x^T \theta^* < 1\}$ have $\alpha_i^* = C/n$.

Role of C

- In the primal, increasing C penalizes errors more (and puts less emphasis on minimizing $\|\theta\|$, that is, maximizing the margin).
- In the dual, decreasing C forces the α_is to be small. So the solution is not strongly influenced by a single outlier.

ν-SVM

An alternative parameterization:

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

that is,

$$\begin{aligned} \min_{\theta,\rho,\xi_i} \quad & \frac{1}{2}\|\theta\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & \rho \geq 0 \\ & \xi_i \geq 0 \\ & y_i\theta' x_i \geq \rho - \xi_i. \end{aligned}$$

ν-SVM

- C is replaced by parameter ν.
- New variable ρ: Points with $\xi_i = 0$ are at distance $\rho/\|\theta\|$ from the decision boundary.
- We can calculate the Lagrangian (with Lagrange multipliers γ, β_i, and α_i, respectively, for the three constraints), hence the dual, as before. We get

$$\theta^* = \sum \alpha_i y_i x_i$$
$$\nu = \sum \alpha_i - \gamma$$
$$\alpha_i + \beta_i = \frac{1}{n}.$$

So we can drop the γ and β_i variables.

ν-SVM

The dual problem is:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & 0 \le \alpha_i \le \frac{1}{n}, \\ & \sum \alpha_i \ge \nu. \end{aligned}$$

ν-SVM

Theorem: If the solution satisfies $\rho > 0$, then

$$
\begin{aligned}
\left|\{i : y_i x_i^T \theta < \rho\}\right| &\overset{(1)}{\leq} \left|\left\{i : \alpha_i = \frac{1}{n}\right\}\right| \\
&\overset{(2)}{\leq} \nu n \\
&\overset{(3)}{\leq} |\{i : \alpha_i > 0\}| \\
&\overset{(4)}{\leq} \left|\{i : y_i x_i^T \theta \leq \rho\}\right| .
\end{aligned}
$$

ν-SVM

Proof:

1. Complementary slackness: $y_i x_i^T \theta < \rho$ implies $\xi_i > 0$ implies $\beta_i = 0$ implies $\alpha_i = 1/n$.

2. Complementary slackness: $\rho > 0$ implies $\gamma = 0$ implies $\sum \alpha_i = \nu$ implies

$$\begin{aligned}\sum_i 1[\alpha_i = 1/n] &= \sum_i n\alpha_i 1[\alpha_i = 1/n] \\ &\leq \sum_i n\alpha_i \\ &= \nu n.\end{aligned}$$

ν-SVM

3. Since $\alpha_i \leq 1/n$, we have

$$\nu \leq \sum \alpha_i \leq \frac{1}{n} \sum 1[\alpha_i > 0].$$

4. Complementary slackness: $\alpha_i > 0$ implies $y_i x_i^T \theta = \rho - \xi_i \leq \rho$.

So ν is a natural parameter. It is approximately the proportion of mistakes. More precisely, it lies between the number of support vectors that fall in the wrong open halfspace, $\{x : x^T \theta^* < 1\}$, and the number of support vectors.

But there is always a suitable choice of C to give the same solution:

Theorem: If the ν-SVM has a solution with $\rho > 0$, then the SVM with $C = 1/\rho$ gives the same decision function.

Representer Theorem

We have seen that, for both the hard margin and soft margin SVM, the optimal θ^* has the form

$$\theta^* = \sum_i \beta_i x_i$$

for some x_i. The representer theorem, which we are about to see, shows that this is always true whenever we solve an optimization problem like

$$\min_\theta \qquad \frac{1}{2}\|\theta\|^2 + L(\theta^T x_1, \ldots, \theta^T x_n),$$

for some L (which corresponds to a surrogate risk).

Representer Theorem

Theorem: Fix a kernel k with corresponding RKHS $\mathcal{H}$. For any (loss) function $L : \mathbb{R}^n \to \mathbb{R}$ and any non-decreasing $\Omega : \mathbb{R} \to \mathbb{R}$, if

$$\begin{aligned} \min_{f \in \mathcal{H}} J(f) &:= \min_{f \in \mathcal{H}} \left(L(f(x_1), \ldots, f(x_n)) + \Omega(\|f\|^2_{\mathcal{H}}) \right) \\ &= J^*, \end{aligned}$$

then for some $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$,

$$f(\cdot) = \sum \alpha_i k(x_i, \cdot)$$

satisfies $J(f) = J^*$. Furthermore, if Ω is increasing, then each minimizer of $J(f)$ can be expressed in this form.

Representer Theorem: Proof

Consider the projection $f_{||}$ on to the subspace

$$\text{span}\{k(x_i, \cdot) : 1 \leq i \leq n\}.$$

Write $f = f_{||} + f_{\perp}$. Then

$$\begin{aligned} f(x_i) &= \langle f, k(x_i, \cdot) \rangle \\ &= \langle f_{||}, k(x_i, \cdot) \rangle \\ &= f_{||}(x_i). \end{aligned}$$

Representer Theorem: Proof

But

$$\|f\|^2 = \|f_{||}\|^2 + \|f_{\perp}\|^2 \geq \|f_{||}\|^2.$$

So

$$\begin{aligned} J(f) &= L(f) + \Omega(\|f\|_{\mathcal{H}}^2) \\ &= L(f_{||}) + \Omega(\|f_{||}\|_{\mathcal{H}}^2 + f_{\perp}\|_{\mathcal{H}}^2) \\ &\geq L(f_{||}) + \Omega(\|f_{||}\|_{\mathcal{H}}^2). \end{aligned}$$

Representer Theorem

That is, we can view an SVM (and any other M-estimator that includes an RKHS norm regularization term in its criterion) as minimizing an objective over all elements of the RKHS, but the solution only needs to be a finite expansion. So we can write an optimization problem like this:

$$\min_{f \in \mathcal{H}} \qquad \frac{1}{2}\|f\|_{\mathcal{H}}^2 + \frac{C}{n}\sum_{i=1}^{n} \ell\left(f(x_i), y_i\right).$$

as:

$$\min_{\beta \in \mathbb{R}^n} \qquad \frac{1}{2}\beta^T K \beta + \frac{C}{n}\sum_{i=1}^{n} \ell\left(\sum_j \beta_j k(x_j, x_i), y_i\right).$$