CS281B/Stat241B. Statistical Learning Theory. Lecture 21.

Peter Bartlett

Overview

- Support vector machines
  - Hard margin
  - Detour into optimization
    (Lagrangian, duality, saddle point, KKT conditions)
  - Dual form of SVM: support vectors
  - Kernels
  - SVM and the convex hull of the data.

Recall: Perceptron convergence theorem

Given *linearly separable data*, that is, $y_i \theta^T x_i > 0$, the perceptron algorithm has risk (also, regret per round) no more than

$$\frac{R^2}{n\gamma^2},$$

where $\gamma = \min_i \theta^T x_i y_i / \|\theta\|$. (PICTURE)

Support Vector Machine

The *support vector machine* optimizes this bound, by maximizing the margin:

$$\max_{\gamma,\theta} \quad \gamma$$

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

Since we only care about the sign for classification, we can, for instance, fix $\|\theta\| = 1/\gamma$ to simplify the problem slightly:

$$\min_{\theta} \quad \|\theta\|$$

$$\text{s.t.} \quad y_i \theta^T x_i \geq 1 \qquad i = 1, 2, \ldots, n.$$

A brief detour into optimization

For the *primal* convex optimization problem

$$
\begin{aligned}
p^* = \min_{x \in \mathbb{R}^n} &\ f_0(x) \\
\text{s.t. } &\ f_i(x) \leq 0, \qquad i = 1, 2, \ldots, m.
\end{aligned}
$$

Introduce Lagrange multipliers (dual variables) $\lambda_1, \ldots, \lambda_m \geq 0$, and define the Lagrangian $L : \mathbb{R}^{n+m} \to \mathbb{R}$ as

$$L(x, \lambda) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x).$$

Dual problem

- The primal problem is the value of the min-max game:

$$p^* = \inf_x \sup_{\lambda \geq 0} L(x, \lambda).$$

(Because for an infeasible x, $L(x, \lambda)$ can be made infinite, and for a feasible x, the $\lambda_i f_i(x)$ terms will become zero.)

- Define the *dual* problem as

$$d^* = \sup_{\lambda \geq 0} g(\lambda) := \sup_{\lambda \geq 0} \inf_x L(x, \lambda).$$

- In a zero sum game, it's always better to choose second:

$$p^* = \inf_x \sup_{\lambda \geq 0} L(x, \lambda) \geq \sup_{\lambda \geq 0} \inf_x L(x, \lambda) = d^*.$$

This is called *weak duality*.

Strong duality

- If there is a *saddle point* (x^*, λ^*), so that for all x and $\lambda \geq 0$,

$$L(x^*, \lambda) \leq L(x^*, \lambda^*) \leq L(x, \lambda^*),$$

then we have *strong duality*:

$$p^* = \inf_x \sup_{\lambda \geq 0} L(x, \lambda) = \sup_{\lambda \geq 0} \inf_x L(x, \lambda) = d^*.$$

This is because:

$$\begin{aligned} \inf_x \sup_{\lambda \geq 0} L(x, \lambda) &\leq \sup_{\lambda \geq 0} L(x^*, \lambda) \\ &= L(x^*, \lambda^*) \\ &= \inf_x L(x, \lambda^*) \\ &\leq \sup_{\lambda \geq 0} \inf_x L(x, \lambda). \end{aligned}$$

Strong duality

There are other sufficient conditions for strong duality (e.g., f_0, f_i convex, and Slater's condition: some x is strictly feasible, that is, satisfies the constraints with strict inequalities).

Complementary slackness

Suppose $p^* = d^*$. Then for primal solution x^*, dual solution λ^*, we have

$$
\begin{aligned}
f_0(x^*) = g(\lambda^*) &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) \right) \\
&\le \left(f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) \right).
\end{aligned}
$$

That is,

$$
\sum_{i=1}^m \lambda_i^* f_i(x^*) \ge 0.
$$

But $\lambda_i^* \ge 0$ and $f_i(x^*) \le 0$, so every term in the sum must be zero:

$$
\lambda_i^* f_i(x^*) = 0.
$$

Complementary slackness

This is known as *complementary slackness*:

if $f_i(x^*) < 0$ then $\lambda_i = 0$.

if $\lambda_i > 0$ then $f_i(x^*) = 0$.

Karush-Kuhn-Tucker optimality conditions

If f_0, f_i are convex and differentiable, then x, λ are optimal and the duality gap is zero iff

1. Primal feasibility: $f_i(x) \leq 0$.
2. Dual feasibility: $\lambda_i \geq 0$.
3. Complementary slackness: $\lambda_i f_i(x) = 0$.
4. Stationarity: $\nabla f_0(x) + \sum_i \lambda_i \nabla f_i(x) = 0$.

Support vector machines

$$\min_{\theta} \quad \frac{1}{2}\|\theta\|^2$$

$$\text{s.t.} \quad y_i\theta^T x_i \geq 1, \qquad i = 1, 2, \ldots, n.$$

$$L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{n} \alpha_i(1 - y_i\theta' x_i)$$

$$g(\alpha) = \inf_{\theta} L(\theta, \alpha)$$

setting $$\theta^* = \sum_{i=1}^{n} \alpha_i y_i x_i,$$

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

Support vector machines

If there is a primal feasible point, we can find a strictly feasible point, so we have strong duality.

Notice that we can express the optimal θ^* in terms of the dual solution, α^*, to

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j \\ \text{s.t.} \quad & \alpha_i \geq 0, \qquad i = 1, 2, \ldots, n. \end{aligned}$$

Support vector machines

Complementary slackness tells us about the role of the α_i:

$$\alpha_i > 0 \text{ implies } y_i \theta^{*\prime} x_i = 1,$$
$$y_i \theta' x_i > 1 \text{ implies } \alpha_i = 0.$$

That is, only the points for which the constraints are tight ("support vectors") appear in the sum defining θ^*. (PICTURE)

Support vector machines

As with the perceptron algorithm, we can express the solution in terms of an arbitrary kernel k:

$$\begin{aligned} f_n(x) &= \operatorname{sign}\left(\langle \theta, \Phi(x) \rangle\right) \\ &= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i \left\langle \Phi(x_i), \Phi(x) \right\rangle\right) \\ &= \operatorname{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x)\right), \end{aligned}$$

where α solves the dual problem

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

Another interpretation

We can write the SVM as an equivalent optimization problem, and the dual leads to an alternative interpretation:

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

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

Another interpretation

$$g(\lambda, \beta) = \inf_{\theta, \gamma} L(\theta, \gamma, \lambda, \beta)$$

setting $\theta^* = \frac{1}{2\beta} \sum_{i=1}^{n} \lambda_i y_i x_i$ and $\sum_{i=1}^{n} \lambda_i = 1$

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

i.e., $\min_{\lambda, \beta} \quad \frac{1}{4\beta} \left\| \sum_i \lambda_i y_i x_i \right\|^2 + \beta$

s.t. $\sum_i \lambda_i = 1, \lambda_i \geq 0, \beta \geq 0.$

Another interpretation

We can find the optimal β and simplify this to

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

And we have that the solution is

$$\theta^* = \frac{\sum_i \lambda_i y_i x_i}{\|\sum_i \lambda_i y_i x_i\|},$$

which is the vector in the direction of the smallest element of

$$\text{co}\left\{y_i x_i : i = 1, \ldots, n\right\}.$$

(PICTURE)