SVM as a Convex Optimization Problem

Leon Gu

CSD, CMU

Convex Optimization

- *Convex set*: the line segment between any two points lies in the set.
- *Convex function*: the line segment between any two points $(x, f(x))$ and $(y, f(y))$ lies on or above the graph of f.
- Convex optimization

$$
\begin{aligned}
\texttt{minimize} \quad & f_0(x) && (1)\\
\texttt{s.t.} \quad & f_i(x) \leq 0 \ \ i = 1, \ldots, m && (2)\\
& h_i(x) = 0 \ \ i = 1, \ldots, p && (3)
\end{aligned}
$$

  - f_0 and f_i convex, h_i linear.
  - convex objective function, convex domain (feasible set).
  - any local optimum is also a global optimum.
- Operations preserve convexity
  - (for convex sets) **intersection**, affine transformation, perspective transformation, ...
  - (for convex functions) nonnegative weighted sum, **maximum and supremum**, composition with affine functions, composition with monotonic convex/concave functions, ...

Optimal Separating Hyperplane

Suppose that our data set $\{x_i, y_i\}_{i=1}^N$ is *linear separable*. Define a hyperplane by

$$\{x : f(x) = \beta^T x + \beta_0 = \beta^T (x - x_0) = 0\} \text{ where } \|\beta\| = 1.$$

- $f(x)$ is the sign distance to the hyperplane.
- we can define a classification rule induced by $f(x)$: $\mathtt{sgn}[\beta^T(x - x_0)]$;

Define the **margin** of $f(x)$ to be the minimal $yf(x)$ through the data

$$C = \min_i y_i f(x_i)$$

A *optimal separating hyperplane* is the hyperplane that maximizes the margin,

$$\max_{\beta, \beta_0, \|\beta\|=1} C, \;\; \mathtt{s.t.} \;\; y_i(\beta^T x_i + \beta_0) \geq C, \;\; i = 1, \ldots, N$$

We can get rid of the norm constraint on β,

$$\frac{1}{\|\beta\|} y_i(\beta^T x_i + \beta_0) \geq C$$

and arbitrarily set $\|\beta\| = 1/C$, then we can rephrase the problem as

$$\min_{\beta,\beta_0} \|\beta\|, \;\; \texttt{s.t.} \;\; y_i(\beta^T x_i + \beta_0) \geq 1, \;\; i = 1, \ldots, N$$

This is a **convex optimization** problem.

Soft Margin SVM

The data is not always perfect. We need to extend optimal separating hyperplane to non-separable cases. The trick is to relax the margin constraints by introducing some "slack" variables.

$$\begin{aligned} \texttt{minimize} \quad & \|\beta\| \ \ \texttt{over} \ \beta, \beta_0 & (4) \\ \texttt{s.t.} \quad & y_i(\beta^T x_i + \beta_0) \geq 1 - \xi_i, \ \ i = 1, \ldots, N & (5) \\ & \xi_i \geq 0; \ \ \sum_{i=1}^{N} \xi_i \leq Z & (6) \end{aligned}$$

- still convex.
- $\xi_i > 1$ – misclassification
  $\xi_i > 0$ – the data is correctly classified but lies in the margin.
- Z is a tuning parameter.

How to solve it? Use Lagrange/duality theory.

Lagrangian Theory

Lagrangian theory characterizes the solution of a constrained optimization problem. Recall the *primal* problem:

$$
\begin{aligned}
\texttt{minimize} \quad & f_0(x) && (7)\\
\texttt{s.t.} \quad & f_i(x) \leq 0 \;\; i = 1, \ldots, m && (8)\\
& h_i(x) = 0 \;\; i = 1, \ldots, p && (9)
\end{aligned}
$$

The stationary points are given by

$$\frac{df_0(x)}{dx} + \sum_{i=1}^{m} \lambda_i \frac{df_i(x)}{dx} + \sum_{i=1}^{p} \nu_i \frac{dh_i(x)}{dx} = 0$$

where λ, ν are free parameters called *Lagrange multipliers*. Accordingly, we define *Lagrangian prime function* (or *Lagrangian*) as

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

We define *Lagrangian dual function* $g(\lambda, \nu)$ as

$$g(\lambda, \nu) = \inf_{x \in \mathcal{X}} L(x, \lambda, \nu).$$

The so-called *Lagrangian dual problem* is the following:

$$\begin{aligned} \texttt{maximize} \qquad & g(\lambda, \nu) && (10) \\ \texttt{s.t.} \qquad & \lambda > 0. && (11) \end{aligned}$$

The *weak duality theorem* says

$$g(\lambda, \nu) \leq f_0(x^*) \quad \text{for all} \quad \lambda \quad \text{and} \quad \nu$$

In other words, maximizing $g(\lambda, \nu)$ over λ and ν produce a bound on $f_0(x^*)$ (Note that $g(\lambda, \nu)$ is piecewise linear and convex). The difference between $g(\lambda^*, \nu^*)$ and $f_0(x^*)$ is called the "duality gap".

K.K.T. Conditions

Slater's Theorem (*Strong Duality Theorem*) says: if the constraint functions are affine, the duality gap is zero.
Then, *K.K.T. conditions* provide **necessary and sufficient** conditions for a point x^* to be an optimum

$$\begin{aligned}
\left.\frac{\partial L(x,\lambda^*,\nu^*)}{\partial x}\right|_{x^*} &= 0 && \texttt{first-order derivative of optimality} \\
\lambda_i^* f_i(x^*) &= 0 && \texttt{complementary slackness conditions} \\
\lambda_i^* &\geq 0 && \texttt{dual constraints} \\
f_i(x^*) &\leq 0 && \texttt{prime constraints} \\
h_i(x^*) &= 0 && \texttt{prime constraints}
\end{aligned}$$

Remarks: complementary slackness conditions are directly related to support vectors.

The Dual Problem

Recall the prime problem (soft-margin SVM)

$$
\begin{aligned}
\texttt{minimize} \quad & \|\beta\| \ \texttt{ over } \beta, \beta_0 && (12) \\
\texttt{s.t.} \quad & y_i(\beta^T x_i + \beta_0) \geq 1 - \xi_i, \ \ i = 1, \ldots, N && (13) \\
& \xi_i \geq 0; \ \sum_{i=1}^{N} \xi_i \leq Z && (14)
\end{aligned}
$$

Obviously strong duality holds. So we can find its dual problem by the following steps

1. Define Lagrange primal function (and Lagrange multipliers).
2. Take the first-order derivatives w.r.t. β, β_0 and ξ_i, and set to zero.
3. Substitute the results into the primal function.

$$\begin{aligned}
\texttt{Maximize} \quad & L_D = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{i'=1}^{N} \alpha_i \alpha_{i'} y_i y_{i'} \langle x_i, x_{i'} \rangle && (15) \\
\texttt{s.t.} \quad & 0 \le \alpha_i \le \gamma, \;\; i = 1, \ldots, N && (16) \\
& \sum_{i=1}^{N} \alpha_i y_i = 0 && (17) \\
\texttt{Solution:} \quad & && (18) \\
& \hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i x_i && (19) \\
& f(x) = \beta^T x + \beta_0 = \sum_{i=1}^{N} \hat{\alpha}_i y_i \langle x_i, x \rangle + \beta_0 && (20)
\end{aligned}$$

- Sparse representation: the separating hyperplane $f(x)$ is spanned those data points i where $\alpha_i \neq 0$, called **Support Vectors**.
  - follows directly from complementary slackness conditions: $\alpha_i \left[y_i(\beta^T x_i + \beta_0) + \xi_i - 1 \right] = 0$
- Both the estimation and the evaluation of $f(x)$ only involve **dot product**.