

How to Solve A Support Vector Machine Problem

Hao Zhang
Colorado State University



Colorado State University
Extension



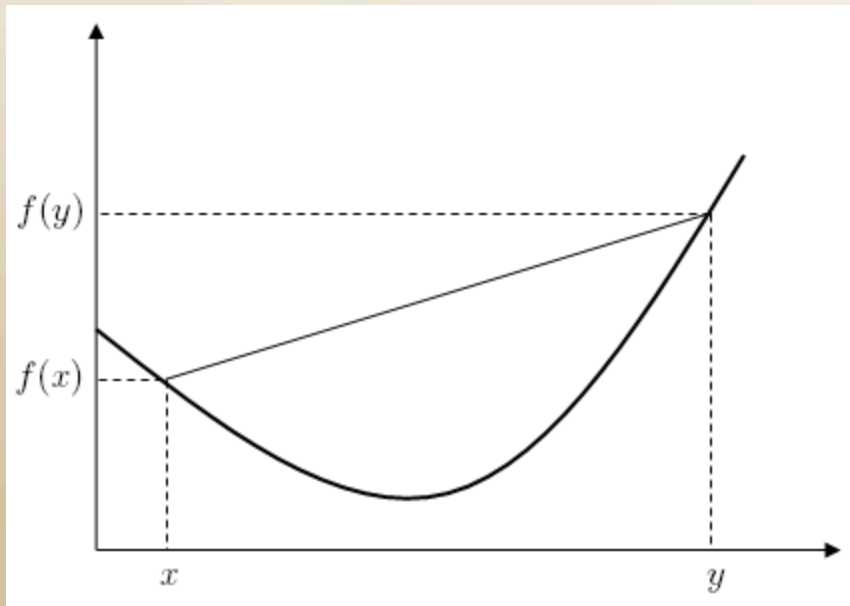
Outline

- Preliminaries
- Support Vector Machine (SVM)
- Reference



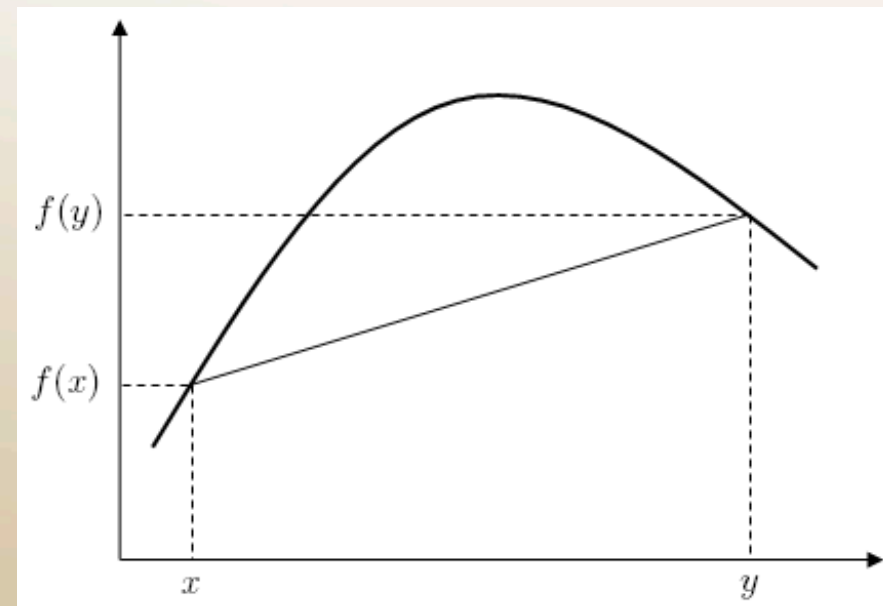
Preliminaries

Convex function



Local minimum is global minimum

Concave function



Local maximum is global maximum



Primal Problem:

$$\text{minimize } f_0(x)$$

$$\text{subject to } f_i(x) \leq 0, i = 1, \dots, m$$

$$h_i(x) = 0, i = 1, \dots, p$$

We denote p^* as the optimal value of this problem

Lagrange function:

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \quad (\lambda_i \geq 0)$$

λ_i and v_i are called Lagrange multipliers



Lagrange dual function

$$\begin{aligned} g(\lambda, v) &= \inf_x L(x, \lambda, v) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \right) \end{aligned}$$

Note: $g(\lambda, v)$ can be $-\infty$ for certain λ and v

Since the dual function $g(\lambda, v)$ is the point-wise infimum of a family of affine functions of (λ, v) , it is concave. (Don't care about the convexity of $f_i(x)$ or $h_i(x)$)



Lagrange dual function

$$\begin{aligned} g(\lambda, v) &= \inf_x L(x, \lambda, v) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \right) \end{aligned}$$

If x is primal feasible ($f_0(x)$ won't approach infinity),
then

$$g(\lambda, v) \leq f_0(x) \quad (\text{quite easy to prove})$$

What does this mean if $g(\lambda, v) > -\infty$ (called dual feasible)?



$$g(\lambda, v) \leq f_0(x)$$

$g(\lambda, v)$ is **one** lower bound on optimal value p^*

Can we do better?



Sure! Just maximize $g(\lambda, v)$:

$$\begin{array}{ll} \text{maximize} & g(\lambda, v) \\ \text{subject to} & \lambda_i \geq 0 \end{array}$$

This is called Lagrange dual problem, associated with its primal problem.

We denote d^* as the optimal value of this dual problem

This is a convex optimization problem, since the objective to be maximized is concave and the constraint is convex. It's still the case even if the primal problem is not convex.



$$g(\lambda, v) \leq f_0(x)$$



$$d^* \leq p^* \text{ (weak duality)}$$

$p^* - d^*$ is defined as optimal duality gap



If the primal problem is convex, we usually (but not always) have strong duality:

$$d^* = p^*$$

One simple constraint qualification is *Slater's condition*:

There exists an $x \in \text{relint}(D)$ such that

$$f_i(x) < 0, i = 1, \dots, m$$

$$h_i(x) = 0, i = 1, \dots, p$$

then we have strong duality

$$* \text{ relint}(D) = \{x \in D: \exists \epsilon > 0, N_\epsilon(x) \cap \text{aff}(D) \subseteq D\}$$



Duality implemented in algorithms

At k th iteration:

compute $g(\lambda^{(k)}, \nu^{(k)})$ and $f_0(x^{(k)})$

$$g(\lambda^{(k)}, \nu^{(k)}) \leq p^* \leq f_0(x^{(k)})$$

stop if the bound is tight enough



Complementary slackness

Suppose that the primal and dual optimal values are attained and equal (strong duality holds). Denote x^* and (λ^*, v^*) as a dual optimal point, i.e.,

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, v^*) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p v_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p v_i^* h_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$



Complementary slackness

Suppose that the primal and dual optimal values are attained and equal (strong duality holds). Denote x^* and (λ^*, v^*) as a dual optimal point, i.e.,

$$\begin{aligned} f_0(x^*) &= g(\lambda^*, v^*) \\ &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p v_i^* h_i(x) \right) \\ &= f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p v_i^* h_i(x^*) \\ &= f_0(x^*) \end{aligned}$$

So, the two inequalities in this chain hold with equality.



Hence,

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

Since each term is non-positive, we conclude that

$$\lambda_i^* f_i(x^*) = 0, \quad i = 1, \dots, m$$

This is called complementary slackness



Karush-Kuhn-Tucker conditions

Assume $f_0, \dots, f_m, h_1, \dots, h_p$ are differentiable, therefore they have open domains. (Don't care about convexity)

$$\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p \nu_i^* \nabla h_i(x^*) = 0$$



Karush-Kuhn-Tucker (KKT) conditions

Put all the conditions and conclusions together, we get
KKT conditions:

$$\begin{aligned}f_i(x^*) &\leq 0, & i = 1, \dots, m \\h_i(x^*) &= 0, & i = 1, \dots, p \\ \lambda_i^* &\geq 0, & i = 1, \dots, m \\ \lambda_i^* f_i(x^*) &= 0, & i = 1, \dots, m\end{aligned}$$

$$\nabla f_0(x^*) + \sum_{i=1}^m \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^p v_i^* \nabla h_i(x^*) = 0$$

Summary: for *any* optimization problem with differentiable objective and constraint functions for which strong duality obtains, any pair of primal and dual optimal points must satisfy the KKT conditions



The KKT conditions play an important role in optimization.

In general, many algorithms for convex optimization are conceived as, or can be interpreted as, methods for solving the KKT conditions.

Example: SVM!

Fletcher proved that for all SVMs, KKT conditions are *necessary and sufficient* for a solution.

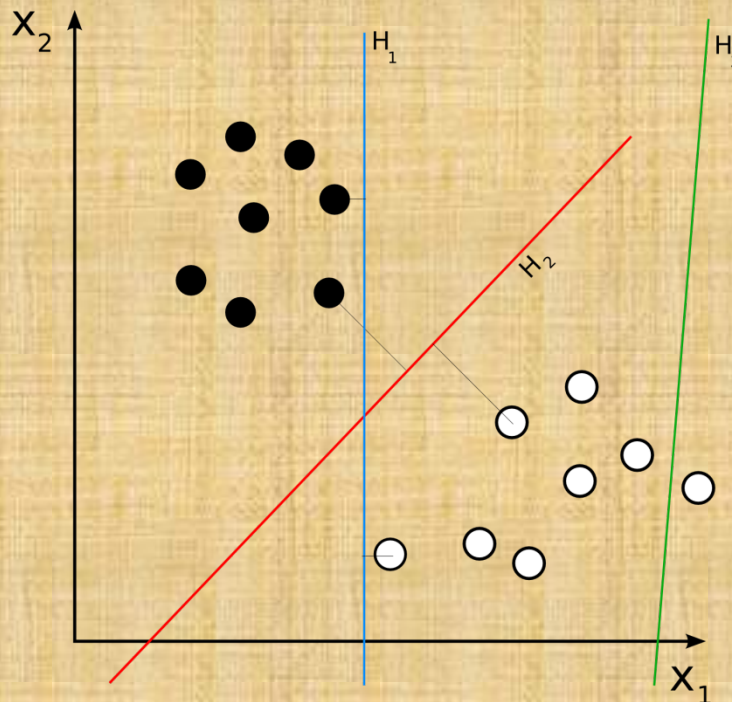
Support Vector Machine (SVM)

Linear SVM

We are given some training data D , a set of n points of the form

$$D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^p, y_i \in \{-1, 1\}\}$$

Our goal is to find the maximum-margin hyperplane that divides the points from these two classes.



H3 (green) doesn't separate the two classes. H1 (blue) does, with a small margin and H2 (red) with the maximum margin.

A hyper plane can be written as

$$w \cdot x - b = 0$$

We want to choose w and b to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible while still separating the data.

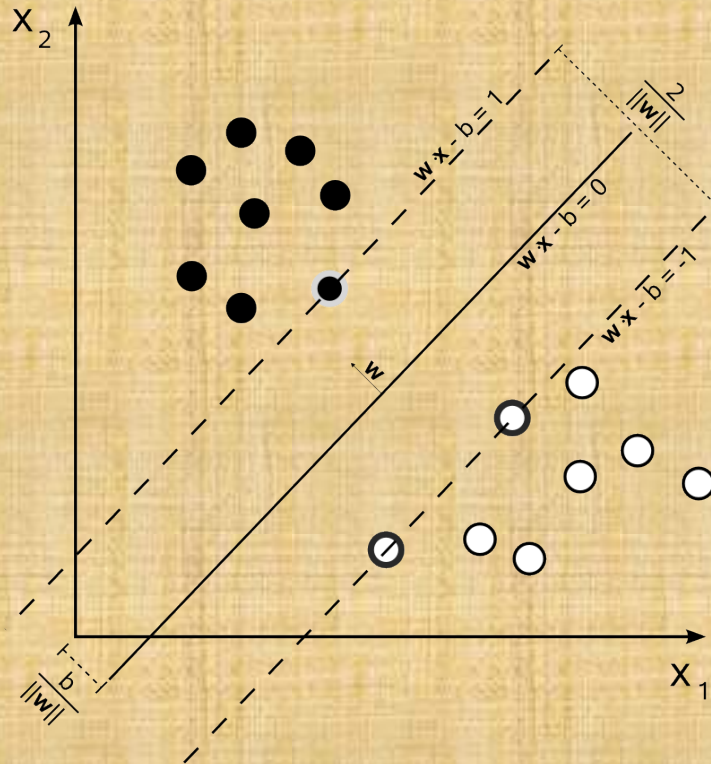
These hyperplanes can be described by the equations:

$$w \cdot x - b = 1$$

and

$$w \cdot x - b = -1$$

(Uniqueness of the parameters)



the distance between
these two hyperplanes
is:

$$\frac{2}{\|w\|}$$

Maximum-margin hyperplane and margins for an SVM trained with samples from two classes. Samples on the margin are called the support vectors.

http://en.wikipedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png

As we also have to prevent data points from falling into the margin, we add the following constraint: for each x_i either

$$w \cdot x_i - b \geq 1 \quad \text{for } y_i = 1$$

or

$$w \cdot x_i - b \leq -1 \quad \text{for } y_i = -1$$

This can be written as

$$y_i(w \cdot x_i - b) \geq 1 \quad \text{for all } i$$

The primal form of this problem is

$$\text{Minimize } \frac{1}{2} ||w||^2$$

$$\text{subject to } y_i(w \cdot x_i - b) \geq 1$$

Its dual form is

$$\max_{\alpha} \min_{w,b} \{L(\alpha, w, b)\}$$

$$\text{where } L(\alpha, w, b) = \frac{1}{2} ||w||^2 - \sum \alpha_i [y_i(w \cdot x_i - b) - 1]$$

$$\text{subject to } \alpha_i \geq 0$$

$$\min_{w,b}\{L(\alpha, w, b)\} \Rightarrow$$

$$\frac{\partial L(\alpha, w, b)}{\partial w} = w - \sum \alpha_i y_i x_i = 0$$

$$\frac{\partial L(\alpha, w, b)}{\partial b} = \sum \alpha_i y_i = 0$$

Geometric view

$$\therefore \min_{w,b}\{L(\alpha, w, b)\} = \sum \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i^T x_j = \tilde{L}(\alpha)$$

Our goal is

$$\max_{\alpha} \tilde{L}(\alpha)$$

Subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$

How to compute w ?

$$w = \sum \alpha_i y_i x_i$$

How to compute b ?

Complementary slackness:


$$\alpha_i [y_i (w \cdot x_i - b) - 1] = 0$$

$$\therefore \text{for those } \alpha_i > 0 \Rightarrow b = w \cdot x_i - y_i$$

(In practice, it's safer to compute the mean)

Kernel SVM

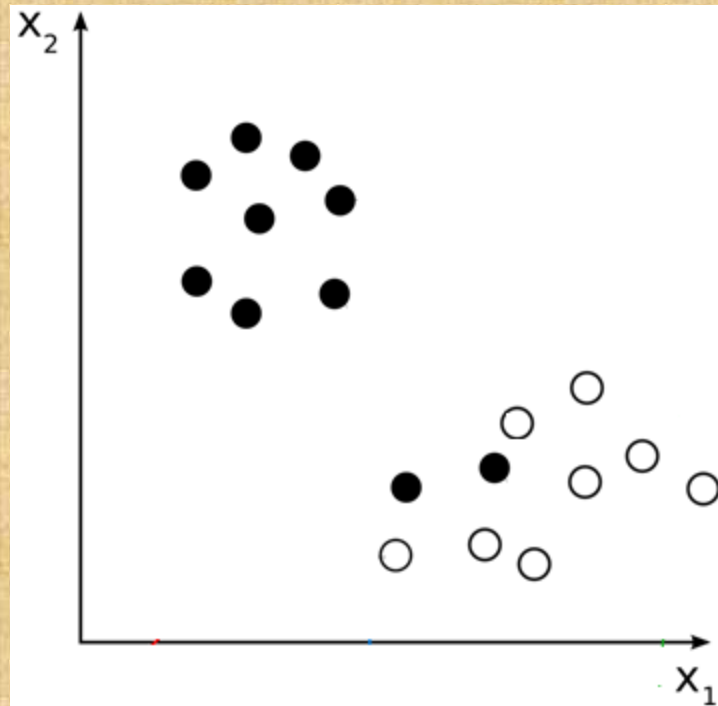
Revisit $\tilde{L}(\alpha)$

$$\tilde{L}(\alpha) = \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \boxed{x_i^T x_j}$$

$$k(x_i, x_j)$$

Some examples of kernels from Wikipedia:

- **Polynomial (homogeneous):** $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d$
- **Polynomial (inhomogeneous):** $k(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^d$
- **Gaussian radial basis function:** $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2)$, for $\gamma > 0$. Sometimes parametrized using $\gamma = 1 / 2\sigma^2$
- **Hyperbolic tangent:** $k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\kappa \mathbf{x}_i \cdot \mathbf{x}_j + c)$, for some (not every) $\kappa > 0$ and $c < 0$

What we have discussed so far focused on separable cases.
what if the data is not separable?



Relax the constraints!

$$y_i(w \cdot x_i - b) \geq 1$$



$$y_i(w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

Primal problem becomes:

$$\min_{w, \xi, b} \left\{ \frac{1}{2} ||w||^2 + C \sum \xi_i \right\}$$

C is a trade-off between a large margin and a small error penalty.

Convert this primal form to dual form. (Same idea!)



Reference

- [1] http://en.wikipedia.org/wiki/Support_vector_machine
- [2] “A Tutorial on Support Vector Machines for Pattern Recognition”,
CHRISTOPHER J.C. BURGESS
- [3] “Convex Optimization”, Stephen Boyd, Lieven Vandenberghe
- [4] “ECEN 629 Lecture 5: Duality and KKT Conditions”, Shuguang Cui
- [5] <http://en.wikipedia.org/wiki/Relint>
- [6] http://en.wikipedia.org/wiki/Affine_hull



Thanks !