Lab 3-2
Lagrangian Multiplier and Duality

Tutorial for Optimization
DMKM Course

Xinyu WANG
wangxinyu.xenia@gmail.com
9-Dec-2013

A brief review

- Types of optimization problems:
  - Linear programming
    - Linear objective function + linear constraints
  - Quadratic programming
    - Quadratic objective function + linear constraints
  - Non linear programming
    - Objective function or some of the constraints are nonlinear

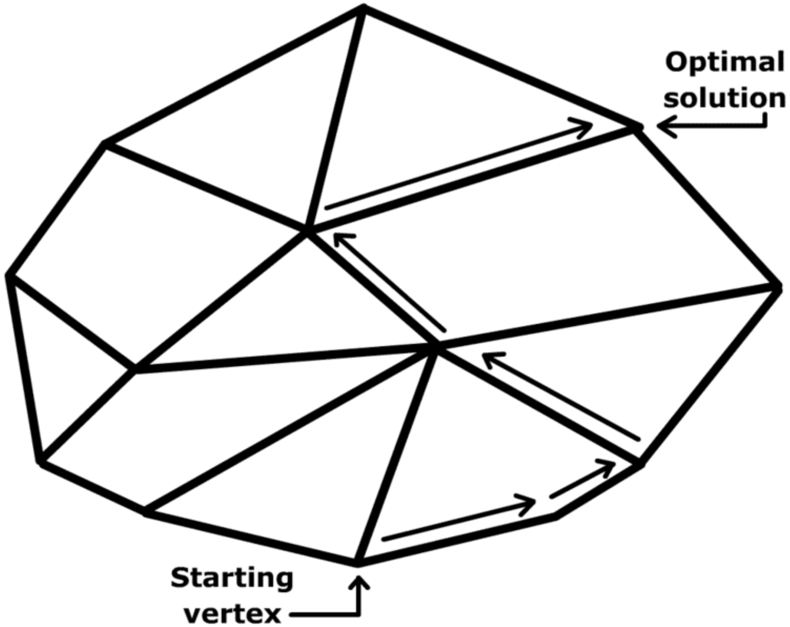
- Types of solutions:
  - Simplex method, for linear programming
  - Lagrange multiplier, for problems with equality constraints
  - KKT conditions, for non linear programming

Simplex method

Standard form:

$$
\begin{aligned}
(LP)\quad \min \quad & c.x \\
\text{s.t.} \quad & Ax = b \\
& x \geq 0
\end{aligned}
$$

To use simplex method, when there is inequality constraint, it is necessary to convert the inequality into equality constraint by adding slack variable or surplus variable.



Actually, in this way, the constraints consists a polytope (in 2-D) . The simplex method begins at a starting vertex and moves along the edges of this polytope until it reaches the vertex of the optimum solution.

The Method of Lagrange Multipliers

This method is used to solve optimization problems with equality constraints. The idea is to reduce constrained optimization into unconstrained.

$$\text{Min } f(x)$$
$$\text{s.t. } \mathrm{h}(x) = b$$

Lagrangian function : $L(x, \boldsymbol{\lambda}) = f(x) - \boldsymbol{\lambda}[\mathrm{h(x) - b}]$

The interesting thing is that: under the constraint, when x makes $f(x)$ to reach its optimal value, it makes $L(x, \boldsymbol{\lambda})$ to reach its optimal value too. So solving Lagrangian function and getting its optimal value equals to solving primal optimization problem.

Solving Lagrangian function is an easier thing to do, for it builds a system of equations:

$$\begin{cases} \dfrac{\partial L(x, \boldsymbol{\lambda})}{\partial x} = 0 \\ \dfrac{\partial L(x, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = 0 \end{cases} \xrightarrow{\text{yields}} \begin{cases} x^* = \\ \boldsymbol{\lambda}^* = \end{cases} \xrightarrow{\text{yields}} L^*(x^*, \boldsymbol{\lambda}^*) = f(x^*) - \boldsymbol{\lambda}^*[\mathrm{h}(x^*) - \mathrm{b}]$$

As $L^*(x^*, \boldsymbol{\lambda}^*) = f^*(x)$, primal problem is solved.

An example of Lagrangian Multipliers

Primal problem: Min $f(x, y) = x^2 + y^2$

s.t. $x + y = 10$

Convert constraint to: $x + y - 10 = 0$

Lagrangian function: $L(x, y, \boldsymbol{\lambda}) = x^2 + y^2 + \boldsymbol{\lambda}(\mathrm{x} + \mathrm{y} - 10)$

The system of equations derived:

$$\begin{cases} \dfrac{\partial L(x, y, \boldsymbol{\lambda})}{\partial x} = 2x + \boldsymbol{\lambda} = 0 \\ \dfrac{\partial L(x, y, \boldsymbol{\lambda})}{\partial y} = 2y + \boldsymbol{\lambda} = 0 \\ \dfrac{\partial L(x, y, \boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = x + y - 10 = 0 \end{cases} \xrightarrow{\text{yields}} \begin{cases} x = 5 \\ y = 5 \\ \boldsymbol{\lambda} = -10 \end{cases} \xrightarrow{\text{yields}} L^*(x^*, y^*, \boldsymbol{\lambda}^*) = 50$$

As $L^*(x^*, y^*, \boldsymbol{\lambda}^*) = f^*(x^*, y^*)$, so the optimal value of primal problem is 50.

Solving equation system by matrix

$$\begin{cases} 2x+\boldsymbol{\lambda}=0 \\ 2y+\boldsymbol{\lambda}=0 \\ x+y-10=0 \end{cases} \quad \mathit{equal\ to} \quad \underbrace{\begin{bmatrix}2&0&1\\0&2&1\\1&1&0\end{bmatrix}}_{\textit{Coefficient matrix}} \begin{bmatrix}x\\y\\\boldsymbol{\lambda}\end{bmatrix} = \underbrace{\begin{bmatrix}0\\0\\10\end{bmatrix}}_{\textit{Variable vector}}$$

Let *C* denote Coefficient matrix, and let *k* denote variable vector:

$$\begin{bmatrix}x\\y\\\boldsymbol{\lambda}\end{bmatrix} = C^{-1}\cdot k = \begin{bmatrix}5\\5\\-10\end{bmatrix}$$

Notations and concepts

$$\min_{x \in R^n} f(x) \quad subject\ to \begin{cases} c_i(x) = 0, i\epsilon\varepsilon \\ c_i(x) > 0, i\epsilon\mathrm{I} \end{cases}$$

Active set:

the active set $A(x)$ at any feasible x consists of the equality constraint indices from ε together with the indices of the inequality constraints i for which $c_i(x) = 0$, that is, $A(x) = \varepsilon \cup \{i \in \mathrm{I} | c_i(x) = 0\}$. At a feasible point x, the inequality constraint $i \in \mathrm{I}$ is said to be active if $c_i(x) = 0$, and inactive if the strict inequality $c_i(x) > 0$ is satisfied.

LICQ:

given the point x and the active set $A(x)$ defined above, we say that the **linear independence constraint qualification (LICQ)** holds if the set of active constraint gradients $\{\nabla C_i(x), i \in A\}$ is linearly independent. In general, if LICQ holds, none of the active constraint gradients can be zero.

Refer to book *Numerical Optimization* by Jorge Nocedal Stephen J. Wright

Karush–Kuhn–Tucker conditions

- 1^{st} order necessary conditions
- For a solution in nonlinear programming to be optimized
- Generalize the method of Lagrangian multiplier, as it allows inequality constraints

A necessary condition for some state of affairs S is a condition that must be satisfied in order for S to obtain; A sufficient condition for some state of affair S is a condition that, if satisfied, guarantees that S obtains.

Suppose that x^* is a local solution of $\min_{x \in R^n} f(x) \quad subject\ to \begin{cases} c_i(x) = 0, i\epsilon\varepsilon \\ c_i(x) > 0, i\epsilon\mathrm{I} \end{cases}$, that the functions f and c_i are continuously differentiable, and that the LICQ holds at x^*. Then there is a Lagrange multiplier vector $\boldsymbol{\lambda}^*$, with components $\boldsymbol{\lambda}_i{}^*$, $i \in \varepsilon \cup \mathrm{I}$, such that the following conditions are satisfied at $(x^*, \boldsymbol{\lambda}^*)$

$$
\begin{aligned}
\nabla_x \mathcal{L}(x^*, \lambda^*) &= 0, \\
c_i(x^*) &= 0, \quad \text{for all } i \in \mathcal{E}, \\
c_i(x^*) &\geq 0, \quad \text{for all } i \in \mathcal{I}, \\
\lambda_i^* &\geq 0, \quad \text{for all } i \in \mathcal{I}, \\
\lambda_i^* c_i(x^*) &= 0, \quad \text{for all } i \in \mathcal{E} \cup \mathcal{I}.
\end{aligned}
$$

complementarity conditions. They imply that either constraint i is active or $\boldsymbol{\lambda}_i$*=0, or possibly both.

An example

Primal problem: For b > 0, Min $f(x,y) = -2\sqrt{x} - \log y$

s.t. $x + y \leq b$ over $x \geq 0, y \geq 0$

Transform constraint to: s.t. $b - x - y \geq 0,\ x \geq 0, y \geq 0$

Lagrangian function:

$$L(x,y,\boldsymbol{\lambda}) = (-2\sqrt{x} - \log y) - \boldsymbol{\lambda}(b - x - y)$$

The system of equations derived:

$$\begin{cases} \dfrac{\partial L(x,y,\boldsymbol{\lambda})}{\partial x} = -\dfrac{1}{\sqrt{x}} - \boldsymbol{\lambda} = 0 \\ \dfrac{\partial L(x,y,\boldsymbol{\lambda})}{\partial y} = -\dfrac{1}{y} - \boldsymbol{\lambda} = 0 \\ \dfrac{\partial L(x,y,\boldsymbol{\lambda})}{\partial \boldsymbol{\lambda}} = x + y - b = 0 \end{cases} \rightarrow \begin{cases} x^* = 1/\boldsymbol{\lambda}^2 \\ y^* = -1/\boldsymbol{\lambda} \\ b = x + y \end{cases}$$

therefore:

$$\mathrm{b} - 1/\boldsymbol{\lambda}^2 + 1/\boldsymbol{\lambda} = \ 0$$

Solve this equation, we get $\begin{cases} \boldsymbol{\lambda}_1 = \frac{-1+\sqrt{1+4b}}{2b} \\ \boldsymbol{\lambda}_2 = \frac{-1-\sqrt{1+4b}}{2b} \end{cases}$

Due to that $\mathrm{b} > 0, \boldsymbol{\lambda}_2 < 0$, but the sign of $\boldsymbol{\lambda}_1$ is not definite.

Consider: $L(x, y, \boldsymbol{\lambda}) = (-2\sqrt{x} - \log y) - \boldsymbol{\lambda}(b - x - y)$, when x, y satisfy the constraint, then $(b - x - y) \geq 0$, so only when $\boldsymbol{\lambda} < 0$, there exists optimal value for the Lagrangian function, otherwise, no solution exists. That's why $\boldsymbol{\lambda}_1$ should be eliminated.

With $\boldsymbol{\lambda}_2$, $\begin{cases} y^* = \frac{-1+\sqrt{1+4b}}{2} \\ x^* = \frac{2b+1-\sqrt{1+4b}}{2} \end{cases}$

Optimal value for the objective function:

$$\text{Min } f(x, y) = -2\left(\frac{2b+1-\sqrt{1+4b}}{2}\right)^{1/2} - \log\left(\frac{-1+\sqrt{1+4b}}{2}\right)$$

Duality

First let's define the original optimization problem as a "primal" problem.
Duality theory shows how an alternative problem (dual problem) can be constructed from the primal problem via Lagrangian function.

In some cases, the dual problem is easier to solve computationally than the original (primal) problem. In other cases, the dual can be used to obtain easily **a lower bound** on the optimal value of the objective for the primal problem.

Consider this is our primal problem:

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \le 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

To solve it, we start by the Lagrangian function:

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

α and β are Lagrangian multipliers.

Duality

Let “P” subscript stands for “primal”. Consider

$$\theta_{\mathcal{P}}(w) = \max_{\alpha,\beta\,:\,\alpha_i \geq 0} \mathcal{L}(w,\alpha,\beta)$$

If $\exists w$, it violates any of the prim

either $g_i(w) > 0$ or $h_i(w) \neq 0$ for some i

Then

$$\theta_P(w) = \max_{\alpha,\beta:\alpha_i \geq 0} \left[f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w) \right] = \infty$$

Therefore

$$\theta_P(w) = \begin{cases} f(w) & \textit{if } w \textit{ satisfies primal constraints} \\ \infty & \textit{otherwise} \end{cases}$$

Thus, θ_P takes the same value as the objective in the primal problem while $\forall w$ satisfy all the constraints. Hence, we can consider the primal problem as:

$$\min_w \theta_P(w) = \min_w \max_{\alpha,\beta:\alpha_i \geq 0} L(w,\alpha,\beta)$$

$P^* = \min_w \theta_P(w)$ is the objective value of primal problem.

Duality

But this problem can also be viewed from another perspective, let’s define:

$$\theta_D(\alpha,\beta) = \min_{w} L(w,\alpha,\beta)$$

$$\theta_P(w) = \underset{\alpha,\beta:\alpha_i\geq 0}{max} L(w,\alpha,\beta)$$

D subscript stands for “dual”. Note that in the definition of θ_P we are maximizing w.r.t α, β, but for θ_D, we are minimizing w.r.t w.

$$\min_{w} \theta_P(w) = \min_{w} \max_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta)$$

The dual optimization problem:

$$\max_{\alpha,\beta:\alpha_i\geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0} \underset{w}{min} L(w,\alpha,\beta)$$

This is exactly the same as the primal problem, except that the order of “max” and “min” are exchanged. Let’s define:

$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \theta_D(\alpha,\beta)$ is the objective value of dual problem.

$P^* = \min_{w} \theta_P(w)$ is the objective value of primal problem.

How are the primal and the dual problems related?

$$d^* = \max_{\alpha,\beta:\alpha_i\geq 0} \underset{w}{min} L(w,\alpha,\beta) \leq \min_{w} \max_{\alpha,\beta:\alpha_i\geq 0} L(w,\alpha,\beta) = P^*$$

You can convince yourself that, $\max \min \leq \min max$ always holds.

Duality

Under some condition:

$$d^* = P^*$$

So that we can solve the dual problem in stead of the primal problem. But under what conditions?

$$\begin{aligned} \min_w \quad & f(w) \\ \text{s.t.} \quad & g_i(w) \leq 0, \quad i = 1, \ldots, k \\ & h_i(w) = 0, \quad i = 1, \ldots, l. \end{aligned}$$

Assume:
(1) f and inequality constaints g are convex
(2) Equality constraints h are linear
(3) inequality constaints g are strictly feasible ≡ $\exists w$ makes $g_i < 0$ for all i

Under these assumption, there must exist w^*, α^*, β^* so that w^* is the solution to the primal problem, and α^*, β^* are the solution to the dual problem. And more over:

$$d^* = P^* = L(w^*, \alpha^*, \beta^*)$$

Duality

Plus, w^*, α^*, β^* satisfy the KKT conditions:

$$\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,n \\
\frac{\partial}{\partial \beta_i}\mathcal{L}(w^*,\alpha^*,\beta^*) &= 0, \quad i=1,\ldots,l \\
\boxed{\alpha_i^* g_i(w^*) = 0, \quad i=1,\ldots,k} \\
g_i(w^*) &\le 0, \quad i=1,\ldots,k \\
\alpha^* &\ge 0, \quad i=1,\ldots,k
\end{aligned}$$

The function marked by orange frame is the KKT complementarity condition. It implies that

if ${\alpha^*}_i > 0$ then $g_i(w^*) = 0$, i.e., $g_i(w^*)$ is active when ${\alpha^*}_i > 0$

When w^*, α^*, β^* satisfy the KKT conditions, they are the solutions to the primal and dual problems, and $d^* = P^*$.

Duality

You may wonder why $\alpha_i \geq 0$? Remember the Lagrangian function:

$$\mathcal{L}(w,\alpha,\beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

Let’s simplify it to:

$$L(w,\alpha,\beta) = f(w) + \alpha g(w) + \beta h(w)$$

Suppose $\forall w$ satisfies the primal constraints, i.e.,

$$g(w) \leq 0\ and\ h(w) = 0$$

So $\beta h(w) = 0$ always holds. For $\alpha g(w)$:

$$\begin{cases} \alpha \geq 0 \rightarrow \alpha g(w)\ \leq 0 \rightarrow L(w,\alpha,\beta) = -\infty \rightarrow maxL(w,\alpha,\beta)\ possible \\ \alpha \leq 0 \rightarrow \alpha g(w)\ \geq 0 \rightarrow L(w,\alpha,\beta) = +\infty \rightarrow maxL(w,\alpha,\beta)\ impossible \end{cases}$$

Especially, when $\alpha \geq 0$, it makes $L(w,\alpha,\beta) \leq f(x)$, it is possible to find lower bound of $f(x)$.

An example with SVM

Let’s take SVM as an example, remember the optimization problem of SVM:

$$\min_{w,b} \frac{1}{2}||w||^2$$

$$s.t.\ \ y^i\left(w'x^i + b\right) \geq 1, i = 1, \dots, m$$

We can write the constraints as:

$$g_i(w) = 1 - \ y^i\left(w'x^i + b\right) \leq 0$$

We have this constraint for each training example. According to the KKT dual complementarity condition:

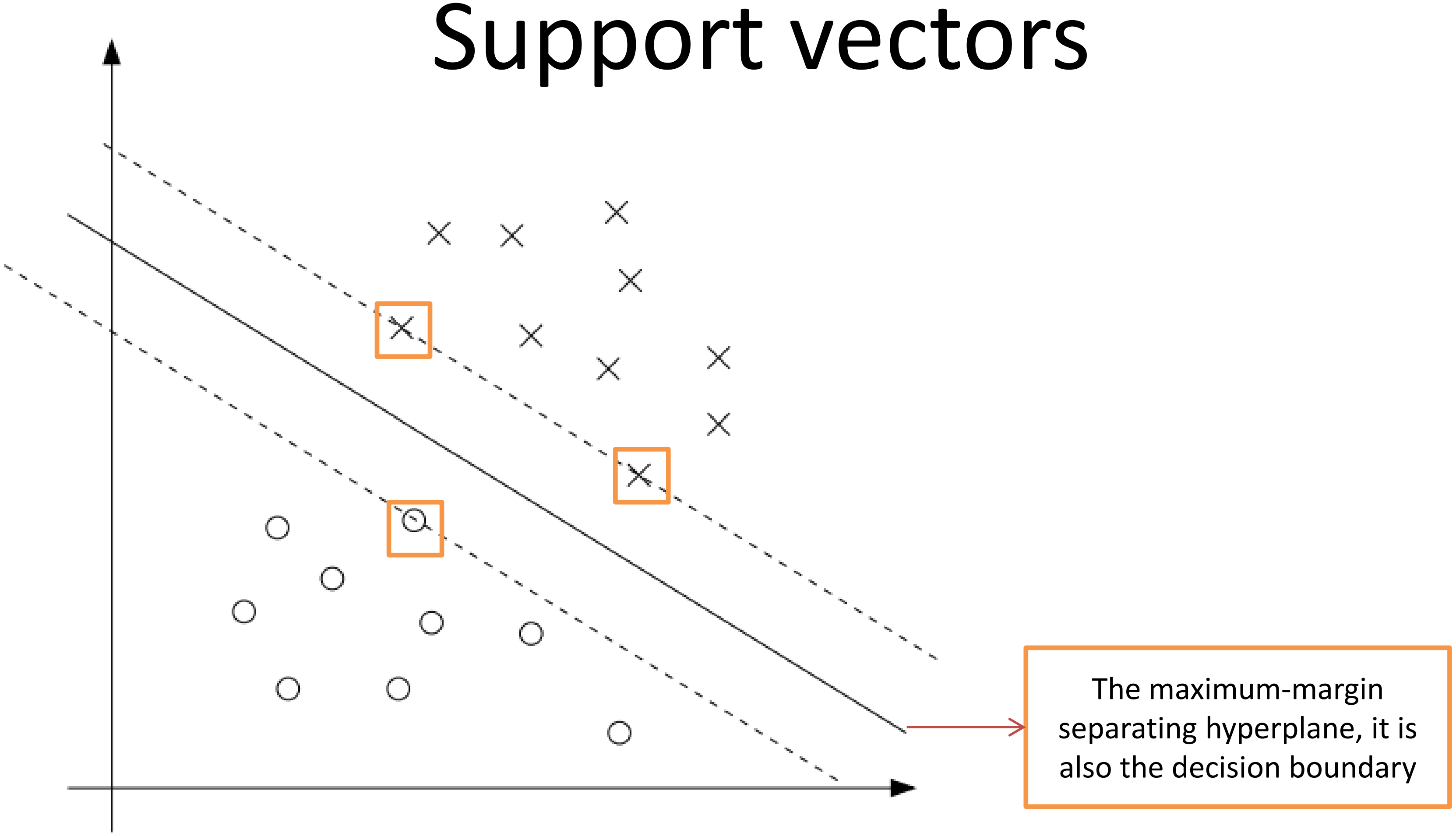
$$\alpha_i^* g_i(w^*) \ = \ 0, \ \ i = 1, \ldots, k$$

When $\alpha_i > 0,\ y^i\left(w'x^i + b\right) = 1$, i.e., points (x^i, y^i) have the smallest functional margin.

Recall the definition of functional margin is:

$$r_f = \ y^i(w'x^i + b)$$

Support vectors



The maximum-margin separating hyperplane, it is also the decision boundary

The three points marked with orange frames have the smallest margins the decision boundary. And only for these three, their α_i is non-zero at the optimal solution to SVM optimization problem.

In SVM, these three points are called the **support vectors.**

Duality with SVM

Lagrangian function for SVM optimization problem:

$$L(w,b,\alpha) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i [\ y^i(w'x^i + b) - 1]$$

Here we have three variables, w and b (from primal problem), α (Lagrange multiplier).
Since we only have inequality constraints here, there is only α but no βLagrange multipliers.

The dual form should take Lagrange multiplier as variable, and be equal to the minimized Lagrangian function w.r.t variables in primal problem, here it should be like:

$$\theta_D(\alpha,\beta) = \min_{w} L(w,\alpha,\beta)$$

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

Thus we need w, b to construct this dual form.
By setting the derivatives of $L(w,b,\alpha)$ w.r.t $w\ and\ b$ to zero, there are:

$$\begin{cases} \dfrac{\partial}{\partial w} L(w,b,\alpha) = w - \displaystyle\sum_{i=1}^{m} \alpha_i y^i x^i = 0 \\ \dfrac{\partial}{\partial b} L(w,b,\alpha) = -\displaystyle\sum_{i=1}^{m} \alpha_i y^i = 0 \end{cases} \equiv \begin{cases} w = \displaystyle\sum_{i=1}^{m} \alpha_i y^i x^i \\ \displaystyle\sum_{i=1}^{m} \alpha_i y^i = 0 \end{cases}$$

Taking w back to the Lagrangian function would get the optimized Lagrangian function.

Use $w = \sum_{i=1}^{m} \alpha_i y^i x^i$ back to the Lagrangian function $L(w, b, \alpha)$, we get:

$$Optimized\ L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j\ \alpha_i \alpha_j (x_i)'^{x_j} - b \sum_{i=1}^{m} \alpha_i y^i$$

$$\sum_{i=1}^{m} \alpha_i y^i = 0$$

Thus:

$$Optimized\ L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j\ \alpha_i \alpha_j (x_i)' x_j$$

Therefore the dual function with Lagrange multiplier as variable:

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

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

So the dual optimization problem:

$$\max_{\alpha,\beta:\alpha_i \geq 0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i \geq 0} \min_{w} L(w,\alpha,\beta)$$

$$\max_{\alpha} \boxed{\left[\sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j\ \alpha_i \alpha_j (x_i)' x_j \right]}$$

s.t. $\alpha_i \geq 0, i = 1, \dots, m$

$$\sum_{i=1}^{m} \alpha_i y^i = 0$$

Let's use $W(\alpha)$ to denote the whole thing

You can verify by yourself that the conditions required for $d^* = P^*$, and the KKT conditions to hold are indeed satisfied in SVM optimization problem. Thus, we can solve the dual instead of solving the primal problem.

Suppose we could solve the dual optimization problem, get α^* and d^*. With α^* known, we can easily get w^*, for $w = \sum_{i=1}^m \alpha_i y^i x^i$.

Having found w^*, by considering the primal problem, it is possible to find the optimal value for the intercept term b as:

$$b^* = -\frac{\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}}{2}.$$

Recall that, in SVM we make prediction on a new input by:

Let's define: $y = w'x + b$ $$y = \begin{cases} 1, & if\ w'x + b \geq 0 \\ -1, & if\ w'x + b < 0 \end{cases}$$

As $w^* = \sum_{i=1}^m \alpha_i y^i x^i$,

$$(w^*)'x + b^* = \left(\sum_{i=1}^m {\alpha^*}_i \, y_i x_i \right)' x + b^* = \sum_{i=1}^m {\alpha^*}_i \, y_i \langle x_i, x \rangle + b^*$$

Recall the training set is $(x_i, y_i), i = 1, \dots, m$

Hence, if we've found the ${\alpha^*}_i$, in order to make a prediction, we just need to calculate the value of the right side, which depends only on the inner product between x and the points in the training set. As we see earlier that most of α_i will all be zero except for the support vectors. Thus many of the terms in the sum above will be zero, and what we need is just the inner products between x and the support vectors (of which there is usually only a small number). The property of SVM enables us to apply Kernels, making itself a very efficient algorithm in earning very high dimensional spaces.

Good books

- Numerical Optimization, by Jorge Nocedal and Stephen J. Wright
- Convex analysis, by R.T. Rockafellar
- Fast Training of Support Vector Machines using Sequential Minimal Optimization, by John C. Platt
- Introduction to Machine Learning, by Ethem Alpaydin