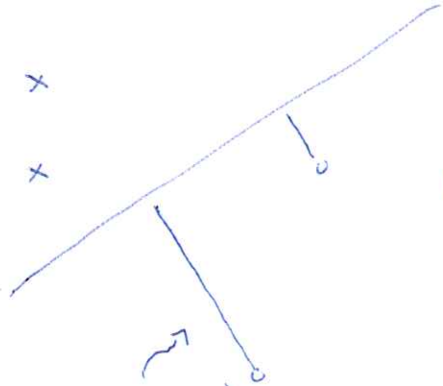
Recall



$$\frac{t_n(w^t x_n + b)}{\|w\|}$$

$\{x : w^t x + b = 0\}$ = hyperplane

$(x_1, t_1) \cdots (x_N, t_N) \quad n = 1 \cdots N \quad t_n \in \{+1, -1\}$

If x_n correctly classified $\left(\operatorname{sgn}(w^t x_n + b) = t_n \right)$

$$\frac{t_n(w^t x_n + b)}{\|w\|} \quad \text{is dist. from} \quad \{x : w^t x + b = 0\}$$

$$\text{Margin} = M(w, b) = \min_n \frac{t_n(w^t x_n + b)}{\|w\|}$$

1 (scaling w)

$$\max_{w,b} \underbrace{\min_n \frac{t_n(w^t x_n + b)}{\|w\|}}_{\text{Margin} = \frac{1}{\|w\|}} \quad \text{s.t.} \quad \underbrace{\min_n t_n(w^t x_n + b) \geq \cancel{0}}_{\text{All points classified correctly.}}$$

$=$

$\vdots$

$$= \min_{w,b} w^t w \quad \text{s.t.} \quad t_n(w^t x_n + b) \geq 1 \quad n = 1 \cdots N$$

(1) $\min w^t w \iff \max \cancel{(w^t w)^{1/2}} \dfrac{1}{\|w\|}$ ~~s.t. $t_n(w^t x_n + b) \geq 1$~~

Interp of w : $\dfrac{1}{\|w\|}$ is margin

(2) Have Q.P. problem.

Want to relax "all correctly classified" assumption

Replace:

$$t_n (w^t x_n + b) \geq 1 \qquad n = 1 \cdots N$$

with

$$t_n (w^t x_n + b) \geq 1 - \underbrace{\xi_n}_{\text{"slack" variables.}} \qquad \xi_n \geq 0 \qquad n = 1 \cdots N$$

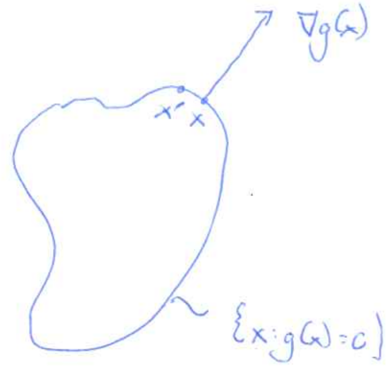
where we try to make $\{ \xi_n \}$ small.

Formulation

$$\min_{\substack{w, b \\ \xi_1 \cdots \xi_N}} w^t w + C \sum_{n=1}^{N} \xi_n \qquad \text{s.t.} \qquad t_n (w^t x_n + b) \geq 1 - \xi_n \quad n = 1 \cdots N$$
$$\xi_n \geq 0 \quad n = 1 \cdots N$$

See penalized_svm.r

Lagrange Multipliers



Have level set $\{x : g(x) = 0\}$

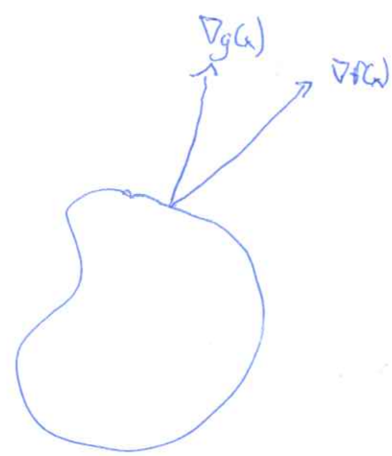
Fact: $\nabla g(x)$ is "normal" (orthogonal)

to $\{x : g(x) = 0\}$.

To see this let $x' \in \{x : g(x) = 0\}$ close to x

$$\underset{0}{\underset{\shortparallel}{g(x')}} \approx \underset{0}{\underset{\shortparallel}{g(x)}} + \underbrace{\nabla g(x) \cdot (x - x')}_{\overset{\shortparallel}{0}}$$

$\Longrightarrow$ $\nabla g(x)$ orthog to $x - x'$



Now consider constrained opt. problem

$$\max_x \; f(x) \quad \text{s.t.} \quad g(x) = 0$$

Solution x must satisfy : $\nabla f(x) = \lambda \nabla g(x)$

($\nabla f(x)$ and $\nabla g(x)$ point in same direction)

Otherwise could "walk along" $\{x : g(x) = 0\}$ making positive proj onto $\nabla f(x)$

thus increasing $f(x)$

The Lagrangian

Consider

$$\max_x f(x) \quad \text{s.t.} \quad g(x) = 0$$

Let

$$L(x, \lambda) = f(x) + \lambda g(x) \qquad \text{(the Lagrangian)}$$

Consider a stationary pt. of Lagrangian
(ie. where $\nabla = 0$)

$$\nabla_{x,\lambda} L(x, \lambda) = 0$$

$$\Longrightarrow$$

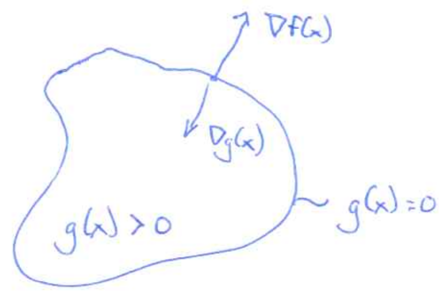
∇_x 1) $\nabla f(x) = -\lambda \nabla g(x)$

∇_λ 2) $g(x) = 0$

Criteria for constrained optimum

$\Longrightarrow$ Stationary pts. of Lagrangian are constrained optima.

Variation



Consider

$$\max_x f(x) \quad \text{s.t.} \quad g(x) \geq 0$$

Sol. must satisfy

① $g(x) > 0$ and $\nabla f(x) = 0$ — Here constraint inactive (corresp. stationary pt. of $L(x,\lambda)$ has $\lambda = 0$)

or

② . ~~$\nabla f(x) = -\lambda \nabla g(x)$~~
$g(x) = 0$ and $\nabla f(x) = -\lambda \nabla g(x)$ — Here constraint active (corresp. st. pt. of $L(x,\lambda)$ has $\lambda > 0$)

Lagrangian Duality

Consider $\max_x f(x)$ s.t. $g_k(x) \geq 0 \quad k = 1 \cdots K$.

$$L(x, \lambda_1 \cdots \lambda_K) = f(x) + \sum_{k=1}^{K} \lambda_k g_k(x)$$

$$\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_K \end{pmatrix} \qquad g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_K(x) \end{pmatrix}$$

$$L(x, \lambda) = f(x) + \lambda^t g(x)$$

Define $\tilde{L}(\lambda) = \max_x L(x, \lambda)$

The <u>dual</u> problem of the original one is

$$\min_\lambda \tilde{L}(\lambda) \quad \text{s.t.} \quad \lambda_1 \cdots \lambda_K \geq 0$$

Dual has interesting connection to original.

For x, λ satisfying constraints, must have

$$\tilde{L}(\lambda) \geq L(x, \lambda) \geq f(x)$$

$$\Longrightarrow \min_{\lambda \geq 0} \tilde{L}(\lambda) \geq \max_{\substack{x \\ g(x) \geq 0}} f(x)$$

However, under some circumstances

$$\min_{\lambda : \lambda \geq 0} \tilde{L}(\lambda) = \max_{x : g(x) \geq 0} f(x)$$

In this case can solve dual to get

$$\lambda^* = \arg\min_{\lambda : \lambda \geq 0} \tilde{L}(\lambda)$$ and get optimal x, x^*, by solving

$$x^* = \arg\min_x L(x, \lambda^*)$$ (unconstrained)

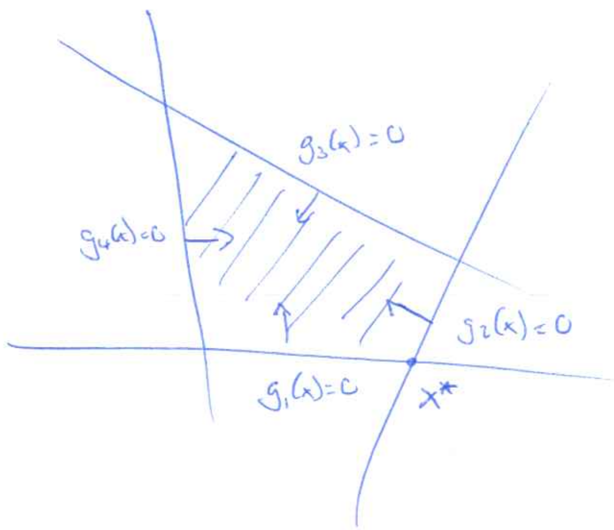
Dual often easier to solve due to simple constraints $\lambda_1 \geq 0 \cdots \lambda_K \geq 0$.

Intergretation of λ^*, x^*

Have $\lambda_1^* \geq 0 \;\ldots\; \lambda_k^* \geq 0$

$$\lambda_k^* = 0 \iff g_k(x^*) > 0 \qquad \text{(constraint inactive)}$$

$$\lambda_k^* > 0 \iff g_k(x^*) = 0 \qquad \text{(constraint active)}$$



Constraints 1,2 active

3,4 inactive

q1

Dual Formulation of SVM

Have

$$\min_{w,b} \ \tfrac{1}{2} w^t w \quad \text{s.t.} \quad t_n (w^t x_n + b) \geq 1 \quad n = 1 \cdots N.$$

$$\Longleftrightarrow$$

$$\max_{w,b} \ -\tfrac{1}{2} w^t w \quad \text{s.t.} \quad t_n (w^t x_n + b) - 1 \geq 0 \quad n = 1 \cdots N.$$

The Lagrangian is

$$L(w, b, \lambda) = -\tfrac{1}{2} w^t w + \sum_{n=1}^{N} \lambda_n \left[t_n (w^t x_n + b) - 1 \right]$$

Dual phrased in terms of

$$\tilde{L}(\lambda) = \max_{w,b} L(w, b, \lambda)$$

$$\nabla_w L(w, b, \lambda) = 0 \implies w = \sum_{n=1}^{N} \lambda_n t_n x_n$$

$$\nabla_b L(w, b, \lambda) = 0 \implies \sum_{n=1}^{N} \lambda_n t_n = 0$$

Get $\tilde{L}(\lambda)$ by substituting these into $L(w, b, \lambda)$

$$\tilde{L}(\lambda) = -\tfrac{1}{2} \sum_{n,m=1}^{N} \lambda_n \lambda_m t_n t_m x_n^t x_m$$

$$+ \cancel{\sum_{n=1}^{N} \lambda_n} \sum_{n,m=1}^{N} \lambda_n \lambda_m t_n t_m x_n^t x_m + \cancelto{0}{\sum_n \lambda_n t_n b} - \sum_n \lambda_n$$

$$= \tfrac{1}{2} \sum_{n,m=1}^{N} \lambda_n \lambda_m t_n t_m x_n^t x_m - \sum_{n=1}^{N} \lambda_n$$

~~Lagrangian Duality~~

~~Have~~ $\cancel{\max_x f(x) \quad \text{s.t.} \quad g_k(x) = 0 \quad k=1\ldots K}$

$$\cancel{L(x, \lambda_1 \cdots \lambda_K) = L(x,\lambda) = f(x) + \sum_{k=1}^{K} \lambda_k g_k(x)}$$

Lagrangian Dual

$$\min_{\lambda_1 \cdots \lambda_N} \quad \frac{1}{2} \sum_{n,m=1}^{N} \lambda_n \lambda_m t_n t_m x_n^t x_m - \sum_{n=1}^{N} \lambda_n \qquad \text{s.t.} \quad \lambda_n \geq 0 \quad n=1 \ldots N$$

$$\sum_{n=1}^{N} \lambda_n t_n = 0$$

This is a new QP problem.

Solve to get $\lambda_1 \cdots \lambda_N$. Most λ's are 0.

$\lambda_n = 0 \iff$ corresp constraint $t_n(w^t x_n + b) \geq 1$ inactive (>1)

$\lambda_n > 0 \iff$ " " " active $(=1)$

The x_n s.t. $\lambda_n > 0$ are "support vectors"

Resulting Classifier

Classify by $\operatorname{sgn}(y(x))$ where

$$y(x) = w^t x + b$$

$$= \sum_{n=1}^{N} \lambda_n t_n x_n^t x + b$$

$$= \sum_{n \in S} \lambda_n t_n x_n^t x + b \qquad \text{where } S = \{n : \lambda_n > 0\}$$

Note: classifier only depends on support vectors

To compute b note for $n \in S$

$t_n(w^t x_n + b) = 1$

$\Rightarrow t_n\left(\sum_{m \in S} \lambda_m t_m x_m^t x_n + b\right) = 1$

Solve for b.

Important Observation

Consider dual SVM formulation

$$\min_{\lambda_1 \cdots \lambda_n} \quad \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \lambda_n \lambda_m t_n t_m x_n^t x_m \; - \; \sum_{n=1}^{N} \lambda_n$$

s.t. $\lambda_n \geqslant \quad n = 1 \cdots N$

$$\sum_{n=1}^{N} \lambda_n t_n = 0$$

Suppose we replace x_n by feature vector $\phi(x_n)$

Note that objective fn depends only on inner products between feature vectors $\phi(x_n) \cdot \phi(x_m)$

A Kernel function, $k(x, x')$ is function s.t.

$$k(x,x') = \phi(x) \cdot \phi(x') \qquad \text{for some } \phi(x)$$

Note Only need to know $k(x,x')$ to ~~[illegible]~~ solve dual (don't need features $\phi(x)$)

Easy to construct Kernel fn's. For ex

① $k(x,x') = x \cdot x'$ is Kernel fn

② $k(x,x')$ is Kernel fn $\Longrightarrow$ $c\,k(x,x')$ is Kernel fn

③ " " " $\Longrightarrow$ $e^{k(x,x')}$ " " "

④ " " " $\Longrightarrow$ $f(x)\, k(x,x')\, f(x')$ " " "

For instance

$\|x - x'\| = (x-x', x-x') = \|x\|^2 - 2\, x \cdot x' + \|x'\|^2$

So

$$e^{-\frac{1}{2\sigma^2}\|x-x'\|^2} = \underbrace{e^{-\frac{1}{2\sigma^2}\|x\|^2} \, \underbrace{e^{\frac{1}{\sigma^2} x \cdot x'}}_{k(x,x') \text{ by } 1,2,3} \, e^{-\frac{1}{2\sigma^2}\|x'\|^2}}_{k(x,x') \text{ by } 4}$$

The feature vector assoc. with $k(x,x') = e^{-\frac{1}{2\sigma^2}\|x-x'\|^2}$

is infinite!

To implement SVM with these features

Substitute $k(x,x') = e^{\frac{-1}{2\sigma^2}\|x-x'\|^2}$

for $x \cdot x'$

in QP formulation of dual and solve for $\lambda_1 \cdots \lambda_N$.

The resulting classifier is

$$\begin{aligned}
\text{Class}(x) &= \text{sgn}(y(x)) \\
&= \text{sgn}(w^t x + b) \\
&= \text{sgn}\left(\sum_{n=1}^{N} \lambda_n t_n k(x_n, x) + b(\lambda)\right) \\
&= \text{sgn}\left(\sum_{n \in S} \lambda_n t_n e^{-\frac{1}{2\sigma^2}\|x_n - x\|^2} + b(\lambda)\right)
\end{aligned}$$