Machine Learning
(機器學習)

Lecture 10: Support Vector Machine (1)

Hsuan-Tien Lin (林軒田)

htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)

Roadmap

1. When Can Machines Learn?
2. Why Can Machines Learn?
3. How Can Machines Learn?
4. How Can Machines Learn Better?
5. Embedding Numerous Features: Kernel Models

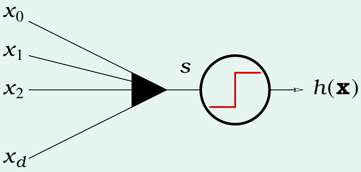
Lecture 10: Support Vector Machine (1)

- Large-Margin Separating Hyperplane
- Standard Large-Margin Problem
- Support Vector Machine
- Motivation of Dual SVM
- Lagrange Dual SVM
- Solving Dual SVM
- Messages behind Dual SVM

Linear Classification Revisited

PLA/pocket

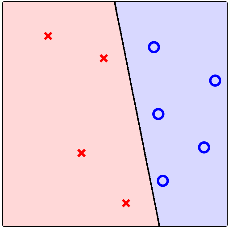
$$h(\mathbf{x}) = \text{sign}(s)$$



plausible err = 0/1

(small flipping noise)

minimize **specially**



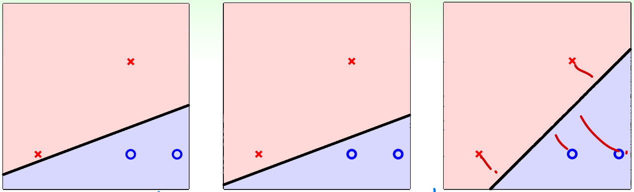
(linear separable)

data是线性可分

linear (hyperplane) classifiers:

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x})$$

Which Line Is Best?



↑答案不是唯一的 那條線好

- PLA? depending on randomness →不一定
- VC bound? whichever you like!

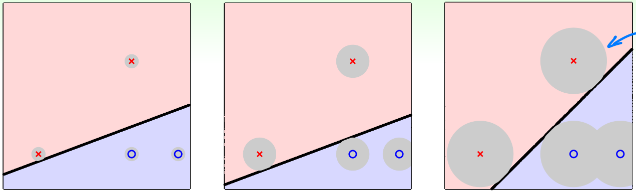
弱

$$E_{\text{out}}(\mathbf{w}) \le \underbrace{E_{\text{in}}(\mathbf{w})}_{0} + \underbrace{\Omega(\mathcal{H})}_{d_{\text{VC}}=d+1}$$

parameter都一樣.

You? **rightmost one, possibly :-)**

Why Rightmost Hyperplane?



noise tolenrence.

informal argument

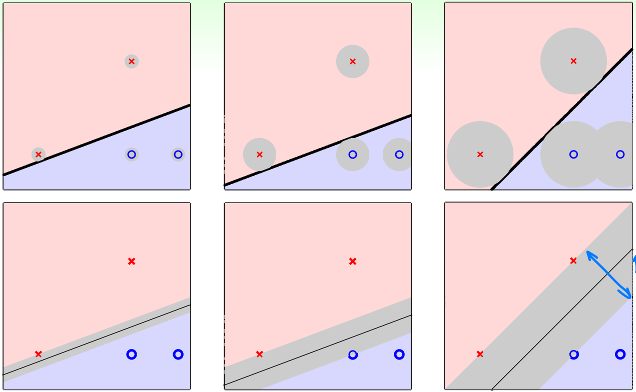
add noise.

if (Gaussian-like) noise on future $\mathbf{x} \approx \mathbf{x}_n$: 原始data.

$\mathbf{x}_n$ further from hyperplane	distance to closest $\mathbf{x}_n$
$\Longleftrightarrow$ tolerate more noise	$\Longleftrightarrow$ amount of noise tolerance
$\Longleftrightarrow$ more robust to overfitting	$\Longleftrightarrow$ robustness of hyperplane

rightmost one: **more robust**
because of **larger distance to closest $\mathbf{x}_n$**

Fat Hyperplane

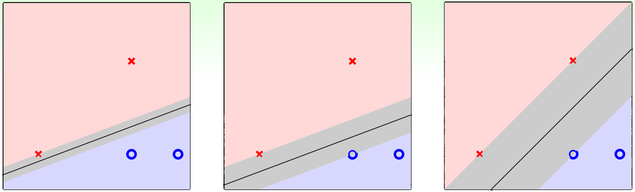


胖线较好.

- **robust** separating hyperplane: **fat**
  —far from both sides of examples
- **robustness** $\equiv$ **fatness**: distance to closest $\mathbf{x}_n$

goal: find fattest separating hyperplane

Large-Margin Separating Hyperplane



$$\max_{\mathbf{w}} \quad \text{fatness}(\mathbf{w})$$

找最胖的线

是線性可分的

subject to $\quad \mathbf{w}$ classifies every $(\mathbf{x}_n, y_n)$ correctly

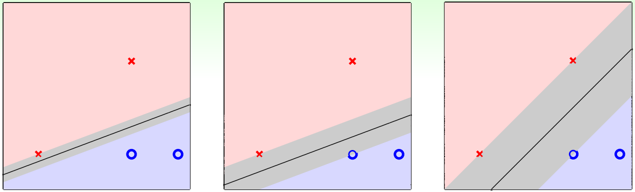
$$\text{fatness}(\mathbf{w}) = \boxed{\min_{n=1,\ldots,N} \text{distance}(\mathbf{x}_n, \mathbf{w})}$$

找最小距離

- fatness: formally called **margin**
- correctness: $y_n = \text{sign}(\mathbf{w}^T \mathbf{x}_n)$ ← 所有in-sample都是对的

goal: find **largest-margin**
separating hyperplane

Large-Margin Separating Hyperplane



$$\max_{\mathbf{w}} \quad \text{margin}(\mathbf{w})$$

subject to every $y_n \mathbf{w}^T \mathbf{x}_n > 0$ ←correctness

$$\text{margin}(\mathbf{w}) = \min_{n=1,\ldots,N} \text{distance}(\mathbf{x}_n, \mathbf{w})$$

- fatness: formally called **margin**
- correctness: $y_n = \text{sign}(\mathbf{w}^T \mathbf{x}_n)$

goal: find **largest-margin** **separating** hyperplane

Questions?

Distance to Hyperplane: Preliminary

$$\max_{\mathbf{w}} \quad \text{margin}(\mathbf{w})$$

subject to every $y_n \mathbf{w}^T \mathbf{x}_n > 0$

$$\text{margin}(\mathbf{w}) = \min_{n=1,\ldots,N} \text{distance}(\mathbf{x}_n, \mathbf{w})$$

截距

'shorten' x and w

distance needs w_0 and $(w_1, \ldots, w_d)$ differently (to be derived)

bias 分开 不把 w_0 放进去

$$b = w_0 \qquad ; \qquad \cancel{x_0 = 1}$$

$$\begin{bmatrix} | \\ \mathbf{w} \\ | \end{bmatrix} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix} ; \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

for this part: $h(\mathbf{x}) = \text{sign}(\mathbf{w}^T \mathbf{x} + b)$

Distance to Hyperplane

want: distance($\mathbf{x}, b, \mathbf{w}$), with hyperplane $\mathbf{w}^T\mathbf{x}' + b = 0$

因為方程式要滿足

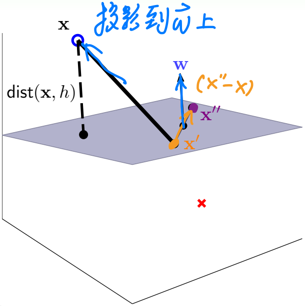
consider $\mathbf{x}'$, $\mathbf{x}''$ on hyperplane

1. $\mathbf{w}^T\mathbf{x}' = -b$, $\mathbf{w}^T\mathbf{x}'' = -b$
2. $\mathbf{w} \perp$ hyperplane:

平面的法向量

$$\left(\mathbf{w}^T \underbrace{(\mathbf{x}'' - \mathbf{x}')}_{\text{vector on hyperplane}} \right) = 0 \qquad -b-(-b)$$

3. distance = project $(\mathbf{x} - \mathbf{x}')$ to $\perp$ hyperplane



$$-\frac{w^T}{\|w\|}x' = \frac{+b}{\|w\|}$$

$$\text{distance}(\mathbf{x}, b, \mathbf{w}) = \underbrace{\left| \frac{\mathbf{w}^T}{\|\mathbf{w}\|}(\mathbf{x} - \mathbf{x}') \right|}_{\text{投影}} \overset{(1)}{=} \frac{1}{\|\mathbf{w}\|} |\mathbf{w}^T\mathbf{x} + b|$$

Distance to Separating Hyperplane

$$\text{distance}(\mathbf{x}, b, \mathbf{w}) = \frac{1}{\|\mathbf{w}\|}|\mathbf{w}^T\mathbf{x} + b|$$

- **separating** hyperplane: for every n

correctness

$$y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0$$

- distance to **separating** hyperplane:

乘上y_n後 abs 可以拿掉

$$\text{distance}(\mathbf{x}_n, b, \mathbf{w}) = \frac{1}{\|\mathbf{w}\|}y_n(\mathbf{w}^T\mathbf{x}_n + b)$$

$$(\alpha_n x_n^2 + b)\frac{1}{\|\Sigma\|}$$

$$\Sigma \alpha_n y_n x_n$$

$$\max_{b,\mathbf{w}} \quad \text{margin}(b, \mathbf{w})$$

subject to every $y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0$

$$\text{margin}(b, \mathbf{w}) = \min_{n=1,\ldots,N} \frac{1}{\|\mathbf{w}\|}y_n(\mathbf{w}^T\mathbf{x}_n + b)$$

Margin of **Special** Separating Hyperplane

$$\max_{b,\mathbf{w}} \quad \text{margin}(b, \mathbf{w})$$

subject to every $y_n(\mathbf{w}^T \mathbf{x}_n + b) > 0$

$$\text{margin}(b, \mathbf{w}) = \min_{n=1,\ldots,N} \frac{1}{\|\mathbf{w}\|} \underbrace{y_n(\mathbf{w}^T \mathbf{x}_n + b)}_{=1}$$

- $\mathbf{w}^T \mathbf{x} + b = 0$ same as $3\mathbf{w}^T \mathbf{x} + 3b = 0$: scaling does not matter
- **special** scaling: only consider separating $(b, \mathbf{w})$ such that

強制scale

$$\min_{n=1,\ldots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1 \Longrightarrow \text{margin}(b, \mathbf{w}) = \frac{1}{\|\mathbf{w}\|} \cdot 1$$

$$\max_{b,\mathbf{w}} \quad \frac{1}{\|\mathbf{w}\|}$$

subject to ~~every $y_n(\mathbf{w}^T \mathbf{x}_n + b) > 0$~~ 消掉了, =1 較嚴格

$$\min_{n=1,\ldots,N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1$$

新condition.

Standard Large-Margin Hyperplane Problem

$$\max_{b,\mathbf{w}} \quad \frac{1}{\|\mathbf{w}\|} \quad \text{subject to} \quad \min_{n=1,\ldots,N} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$

necessary constraints: $y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$ for all n

relax

放鬆條件也沒有損失

original constraint: $\min_{n=1,\ldots,N} \; y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$

want: optimal $(b, \mathbf{w})$ **here (inside)**

反證法：證不可能所有 x_n 都 >1，會取 =1

if optimal $(b, \mathbf{w})$ outside, e.g. $y_n(\mathbf{w}^T\mathbf{x}_n + b) > 1.126$ for all n

—can scale $(b, \mathbf{w})$ to "more optimal" $\left(\frac{b}{1.126}, \frac{\mathbf{w}}{1.126}\right)$ **(contradiction!)**

scale

$\|w\|\downarrow$, $\frac{1}{\|w\|}\uparrow$，這是解更好

(b,w) 不是 optimal：$\rightarrow\leftarrow$.

final change: max $\Longrightarrow$ min, remove $\sqrt{\ }$, add $\frac{1}{2}$

$\frac{1}{\|w\|} \Rightarrow \|w\| \Rightarrow w^Tw \Downarrow \frac{1}{2}w^Tw$

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

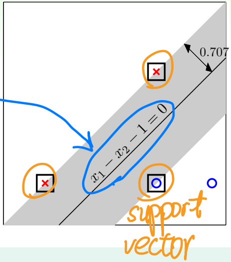
$$\text{subject to} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for all } n$$

Questions?

Support Vector Machine (SVM)

optimal solution: $(w_1 = 1, w_2 = -1, b = -1)$

margin$(b, \mathbf{w})$ $= \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$



- examples on boundary: 'locates' fattest hyperplane
  other examples: **not needed** ← 把這些点去掉也不影响
- call boundary example **support vector** (candidate)

support vector machine (SVM):
learn **fattest hyperplanes**
(with help of **support vectors**)

Solving General SVM

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{subject to} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \text{ for all } n$$

- **not easy manually, of course :-)**
- gradient descent? **not easy with constraints**
- luckily: 限制式:W的二次function
  - (convex) quadratic objective function of $(b, \mathbf{w})$
  - linear constraints of $(b, \mathbf{w})$

—**quadratic programming**

quadratic programming (QP):
'easy' optimization problem

Quadratic Programming

optimal $(b, \mathbf{w}) = ?$

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

subject to $\quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1,$

for $n = 1, 2, \ldots, N$

optimal $\mathbf{u} \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$

u的二次函數　　一次項係數

$$\min_{\mathbf{u}} \quad \underbrace{\frac{1}{2}\mathbf{u}^T\mathrm{Q}\mathbf{u} + \mathbf{p}^T\mathbf{u}}$$

2次項係數

subject to $\quad \mathbf{a}_m^T\mathbf{u} \geq c_m,$

for $m = 1, 2, \ldots, M$

欲求的 variable.　　一次式限制

沒有一次項

objective function: $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$; $\mathrm{Q} = \begin{array}{c|cc} & [b & \mathbf{w}] \\ \hline b & 0 & \mathbf{0}_d^T \\ \mathbf{w} & \mathbf{0}_d & \mathrm{I}_d \end{array}$; $\mathbf{p} = \mathbf{0}_{d+1}$

constraints: $\mathbf{a}_n^T = y_n \begin{bmatrix} 1 & \mathbf{x}_n^T \end{bmatrix}$; $c_n = 1$; $M = N$

SVM with general QP solver:
easy **if you've read the manual :-)**

wTF

SVM with QP Solver

Linear Hard-Margin SVM Algorithm

1. $Q = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathrm{I}_d \end{bmatrix}$; $\mathbf{p} = \mathbf{0}_{d+1}$; $\mathbf{a}_n^T = y_n \begin{bmatrix} 1 & \mathbf{x}_n^T \end{bmatrix}$; $c_n = 1$
2. $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \mathrm{QP}(Q, \mathbf{p}, A, \mathbf{c})$
3. return b & $\mathbf{w}$ as g_{SVM}

坚持所有点都要分开

- **hard-margin**: nothing violate 'fat boundary'
- **linear**: $\mathbf{x}_n$ ← 找分隔线

want **non-linear**?

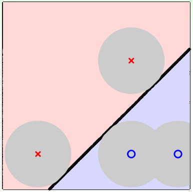
$\mathbf{z}_n = \Phi(\mathbf{x}_n)$—**remember? :-)**

Why Large-Margin Hyperplane?

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{subject to} \quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1 \text{ for all } n$$

転成z空間



对調

一体2面

	minimize	constraint
regularization	E_{in}	$\mathbf{w}^T\mathbf{w} \leq C$
SVM	$\mathbf{w}^T\mathbf{w}$	$E_{\text{in}} = 0$ [and more]

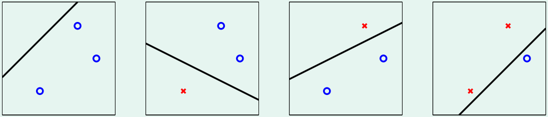
SVM (large-margin hyperplane):
'weight-decay regularization' within $E_{\text{in}} = 0$

Large-Margin Restricts Dichotomies

consider 'large-margin algorithm' $\mathcal{A}_\rho$:
either **returns g with margin$(g) \geq \rho$ (if exists)**, or 0 otherwise

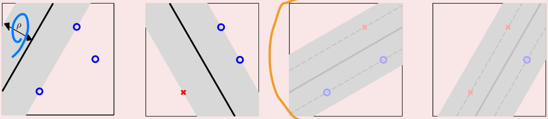
$\mathcal{A}_0$: like PLA $\Longrightarrow$ shatter 'general' 3 inputs

3 input 能 shatter



$\mathcal{A}_{1.126}$: more strict than SVM $\Longrightarrow$ cannot shatter any 3 inputs

但多了限制 ρ寬度



找到ρ=11.26 的 "linear"

fewer dichotomies $\Longrightarrow$ smaller 'VC dim.' $\Longrightarrow$ **better generalization**

VC Dimension of Large-Margin Algorithm

fewer dichotomies $\Longrightarrow$ smaller **'VC dim.'**

considers $d_{\text{VC}}(\mathcal{A}_\rho)$ [data-dependent, need more than VC]

instead of $d_{\text{VC}}(\mathcal{H})$ [data-independent, covered by VC]

generally, when $\mathcal{X}$ in radius-R hyperball:

比原本小

$$d_{\text{VC}}(\mathcal{A}_\rho) \le \min\left(\frac{R^2}{\rho^2}, d\right) + 1 \le \underbrace{d+1}_{d_{\text{VC}}(\text{perceptrons})}$$

ρ是限制的宽度

Benefits of Large-Margin Hyperplanes

	large-margin hyperplanes	hyperplanes	hyperplanes + feature transform Φ
#	even fewer	**not many**	many
boundary	simple	simple	**sophisticated**

- **not many** good, for d_{VC} and generalization
- **sophisticated** good, for possibly better E_{in}

a new possibility: non-linear SVM

	large-margin ✓ **hyperplanes + numerous feature transform Φ** ✓
#	**not many**
boundary	**sophisticated**

ŹŹ win

Questions?

Non-Linear Support Vector Machine Revisited

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

做了転換

$$\text{s. t.} \quad y_n(\mathbf{w}^T \underbrace{\mathbf{z}_n}_{\Phi(\mathbf{x}_n)} + b) \geq 1,$$

$$\text{for } n = 1, 2, \ldots, N$$

Non-Linear Hard-Margin SVM

1. $\mathrm{Q} = \begin{bmatrix} 0 & \mathbf{0}_{\tilde{d}}^T \\ \mathbf{0}_{\tilde{d}} & \mathrm{I}_{\tilde{d}} \end{bmatrix}$; $\mathbf{p} = \mathbf{0}_{\tilde{d}+1}$; $\mathbf{a}_n^T = y_n \begin{bmatrix} 1 & \mathbf{z}_n^T \end{bmatrix}$; $c_n = 1$
2. $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$
3. return $b \in \mathbb{R}$ & $\mathbf{w} \in \mathbb{R}^{\tilde{d}}$ with $g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\Phi(\mathbf{x}) + b)$

控制模型複雜度

- demanded: **not many** (large-margin), but **sophisticated** boundary (feature transform)
- QP with $\tilde{d} + 1$ variables and N constraints —challenging if $\tilde{d}$ large, **or infinite?! :-)**

W dimension 很大?

goal: SVM **without dependence on $\tilde{d}$**

Todo: SVM 'without' $\tilde{d}$

Original SVM

(convex) QP of

- $\tilde{d} + 1$ variables
- N constraints

'Equivalent' SVM

(convex) QP of

- N variables
- $N + 1$ constraints

只跟N有关
N= # of data.

Warning: Heavy Math!!!!!!

- introduce some necessary math without rigor to help **understand SVM deeper**
- **'claim' some results** if details unnecessary
  —like how we 'claimed' Hoeffding

对偶問題

'Equivalent' SVM: based on some **dual problem** of Original SVM

Key Tool: Lagrange Multipliers

乘上係數的 constraint

Regularization by Constrained-Minimizing E_{in}	$\Leftrightarrow$	Regularization by Minimizing E_{aug}
$\min_{\mathbf{w}} E_{\text{in}}(\mathbf{w})$ s.t. $\mathbf{w}^T\mathbf{w} \le C$	rid of constraint	$\min_{\mathbf{w}} E_{\text{aug}}(\mathbf{w}) = E_{\text{in}}(\mathbf{w}) + \frac{\lambda}{N}\mathbf{w}^T\mathbf{w}$

W長度不能太大

- C equivalent to some $\lambda \geq 0$ by checking **optimality condition**

$$\nabla E_{\text{in}}(\mathbf{w}) + \frac{2\lambda}{N}\mathbf{w} = \mathbf{0}$$

- regularization: view λ as **given parameter instead of C**, and solve 'easily'
- dual SVM: view λ's as unknown given the constraints, and **solve them as variables instead** 把λ當變數

how many λ's as variables? SVM有N個constraint.

N—one per constraint

Starting Point: Constrained to 'Unconstrained'

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$-(w^T z_n + b) \geq \rho_-$

$$\text{s.t.} \quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1, \quad \text{for } n = 1, 2, \ldots, N$$

$\Rightarrow$

Lagrange Function

with Lagrange multipliers ~~λ_n~~ α_n,

係數

$$\mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) = \underbrace{\frac{1}{2}\mathbf{w}^T\mathbf{w}}_{\text{objective}} + \sum_{n=1}^{N} \alpha_n \underbrace{(1 - \cancel{y_n}(\mathbf{w}^T\mathbf{z}_n + b))}_{\text{constraint}}$$

1

$+\sum \alpha_n(\rho_- + w^T z_n + b)$

Claim

沒有constraint的最佳化

$$\text{SVM} \equiv \min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) = \min_{b,\mathbf{w}} \left(\infty \text{ if violate} ; \frac{1}{2}\mathbf{w}^T\mathbf{w} \text{ if feasible} \right)$$

調α s.t $\mathcal{L}(\ )$最大

- any 'violating' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \left(\square + \sum_n \alpha_n (\text{some positive}) \right) \to \infty$ — α趕到∞

  壞(b,w)得到∞

- any 'feasible' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \left(\square + \sum_n \alpha_n (\text{all non-positive}) \right) = \square$

  optimal. value.

使得式子看起來沒constraint

constraints now **hidden in** max

Questions?

Strong Duality of Quadratic Programming

min, max 做交換

$$\underbrace{\min_{b,\mathbf{w}}\left(\max_{\text{all }\alpha_n\ge 0}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)}_{\text{equiv. to original (primal) SVM}} = \underbrace{\max_{\text{all }\alpha_n\ge 0}\left(\min_{b,\mathbf{w}}\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})\right)}_{\text{Lagrange dual}}$$

- '=': **strong duality**, true for $\boxed{\text{QP}}$ if
  - convex primal
  - feasible primal (true if Φ-separable) 預
  - linear constraints, SVM 本來就線性

—called constraint qualification ⇒ 能直接解 RHS

exists primal-dual optimal solution $(b, \mathbf{w}, \boldsymbol{\alpha})$ for both sides

Solving Lagrange Dual: Simplifications (1/2)

RHS

$$\max_{\text{all } \alpha_n \geq 0} \left(\overbrace{\underbrace{\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n + b))}_{\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}}^{\text{沒有條件了}} \right)$$

有條件

- inner problem 'unconstrained', at optimal:
  $\frac{\partial \mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial b} = 0 = -\sum_{n=1}^{N} \alpha_n y_n$ 裡面微分=0
- no loss of optimality if solving with constraint $\sum_{n=1}^{N} \alpha_n y_n = 0$

but wait, b **can be removed**

只考慮微分=0的α.

变成常数.

$$\max_{\text{all } \alpha_n \geq 0, \boxed{\sum y_n \alpha_n = 0}} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n)) - \cancel{\sum_{n=1}^{N} \alpha_n y_n \cdot b} \right)$$

Solving Lagrange Dual: Simplifications (2/2)

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0} \left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n)) \right)$$ 處理 W

- inner problem 'unconstrained', at optimal:
  $\frac{\partial \mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})}{\partial w_i} = 0 = w_i - \sum_{n=1}^{N} \alpha_n y_n z_{n,i}$
- no loss of optimality if solving with constraint $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n$

極值發生的位置

but wait!

只找極值處.

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} \left(\boxed{\min_{b,\mathbf{w}}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n - \mathbf{w}^T\mathbf{w} \right)$$

min 也不用做了

$$\Longleftrightarrow \max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\| \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n \right\|^2 + \sum_{n=1}^{N} \alpha_n$$

① ②求b ③求w

3 步 condition

KKT Optimality Conditions

人名

$\frac{1}{\|w\|}$ = distance.

幾乎只有α

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} \; -\frac{1}{2}\Big\|\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n\Big\|^2 + \sum_{n=1}^{N} \alpha_n$$

$// \|w\|^2$

大於≥ 0

$-\frac{1}{2}\|w\|^2 +$

$-\frac{1}{2}\left(-\|\Sigma\|^2 + 2\Sigma\alpha\right) = \frac{1}{\|w\|}$

if primal-dual optimal $(b, \mathbf{w}, \boldsymbol{\alpha})$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$ dual condition.
- dual-inner optimal: $\sum y_n \alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$ ← min的微分
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

↙ 2个至少有1个等於零.

$\frac{\alpha_n}{w^T z_n + b} = 0$

$$\alpha_n\big(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)\big) = 0$$

$y_n w^T z_n + y_n b = 1$

—called **Karush-Kuhn-Tucker (KKT) conditions**, necessary for optimality [& sufficient here]

will use **KKT** to 'solve' $(b, \mathbf{w})$ from optimal $\boldsymbol{\alpha}$

Questions?

Dual Formulation of Support Vector Machine

希望求min：取负

$$\max_{\text{all } \alpha_n \geq 0, \sum y_n \alpha_n = 0, \mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\|\sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n\right\|^2 + \sum_{n=1}^{N} \alpha_n$$

standard hard-margin SVM dual

平方项展开

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \mathbf{z}_n^T \mathbf{z}_m - \sum_{n=1}^{N} \alpha_n$$

subject to $\sum_{n=1}^{N} y_n \alpha_n = 0;$ ← 1个 constraint.

$\alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N$ ← N个 condition

N个α.

(convex) QP of N variables & $N+1$ constraints, as promised

how to solve? **yeah, we know QP! :-)**

Dual SVM with QP Solver

解2次

optimal $\boldsymbol{\alpha} = ?$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} \alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} y_n\alpha_n = 0;$$

$$\alpha_n \geq 0, \quad \text{for } n = 1, 2, \ldots, N$$

表示成QP

optimal $\boldsymbol{\alpha} \leftarrow \text{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T\mathrm{Q}\boldsymbol{\alpha} + \mathbf{p}^T\boldsymbol{\alpha}$$

$$\text{subject to} \quad \mathbf{a}_i^T\boldsymbol{\alpha} \geq c_i, \quad \text{for } i = 1, 2, \ldots$$

- $q_{n,m} = y_n y_m \mathbf{z}_n^T\mathbf{z}_m$ 2次項係數
- $\mathbf{p} = -\mathbf{1}_N$ 1次項係數
- $\mathbf{a}_{\geq} = \mathbf{y}$, $\mathbf{a}_{\leq} = -\mathbf{y}$; $\mathbf{a}_n^T = n$-th unit direction
- $c_{\geq} = 0$, $c_{\leq} = 0$; $c_n = 0$

特殊的bound

note: many solvers treat **equality ($\mathbf{a}_{\geq}, \mathbf{a}_{\leq}$)** & **bound ($\mathbf{a}_n$)** constraints specially for **numerical stability**

Dual SVM with Special QP Solver

dual

optimal $\boldsymbol{\alpha} \leftarrow \text{QP}(\mathrm{Q_D}, \mathbf{p}, \mathrm{A}, \mathbf{c})$

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q_D} \boldsymbol{\alpha} + \mathbf{p}^T \boldsymbol{\alpha}$$

subject to special equality and bound constraints

- $q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m$, often non-zero 不稀疏
- if $N = 30,000$, dense $\mathrm{Q_D}$ (N by N symmetric) takes > 3G RAM
  3万笔，存一半
- need **special solver** for
  - not storing whole $\mathrm{Q_D}$ ← 特殊加速
  - utilizing special constraints properly

  to scale up to large N

usually better to use **special solver** in practice

Optimal $(b, \mathbf{w})$ 有α後,解得(b,w)

KKT conditions

if primal-dual optimal $(b, \mathbf{w}, \boldsymbol{\alpha})$, 中間產物

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n \alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0 \text{ (complementary slackness)}$$

- optimal $\boldsymbol{\alpha} \Longrightarrow$ optimal $\mathbf{w}$? easy above!
- optimal $\boldsymbol{\alpha} \Longrightarrow$ optimal b? a range from primal feasible & equality from **comp. slackness** if one $\alpha_n > 0 \Rightarrow b = y_n - \mathbf{w}^T\mathbf{z}_n$

其中有一人=零

從一個不是零的α_n求得b.

comp. slackness:

$\alpha_n > 0 \Rightarrow$ **on fat boundary (SV!)**

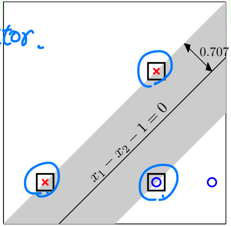
該點剛好是support vector

Questions?

Support Vectors Revisited

- on boundary: 'locates' fattest hyperplane; others: **not needed**
- examples with $\alpha_n > 0$: on boundary

  $\alpha_n > 0$, 該点為support vector.
- call $\alpha_n > 0$ examples $(\mathbf{z}_n, y_n)$ **support vectors** ~~(candidates)~~
- SV (positive α_n) $\subseteq$ SV candidates (on boundary)



α 的 linear combine.

只算SV 的α.

- only SV needed to compute $\mathbf{w}$: $\mathbf{w} = \sum_{n=1}^{N} \alpha_n y_n \mathbf{z}_n = \boxed{\sum_{\text{SV}} \alpha_n y_n \mathbf{z}_n}$
- only SV needed to compute b: $b = y_n - \mathbf{w}^T \mathbf{z}_n$ with any SV $(\mathbf{z}_n, y_n)$

任一个SV.

SVM: learn **fattest hyperplane** by identifying **support vectors** with **dual** optimal solution

① 找SV ② 用SV找optimal.

Summary: Two Forms of Hard-Margin SVM

Primal Hard-Margin SVM

原始的

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

sub. to $\quad y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1,$

for $n = 1, 2, \ldots, N$

- $\tilde{d}+1$ variables,
  N constraints
  —suitable when $\tilde{d}+1$ small
- physical meaning: locate **specially-scaled** $(b, \mathbf{w})$

Dual Hard-Margin SVM

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q_D} \boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha}$$

s.t. $\quad \mathbf{y}^T\boldsymbol{\alpha} = 0;$

$\alpha_n \geq 0$ for $n = 1, \ldots, N$

- N variables,
  $N+1$ simple constraints
  —suitable when N small
- physical meaning: locate **SVs** $(\mathbf{z}_n, y_n)$ & their α_n

both eventually result in optimal $(b, \mathbf{w})$ for fattest hyperplane

$$g_{\text{SVM}}(\mathbf{x}) = \text{sign}(\mathbf{w}^T\Phi(\mathbf{x}) + b)$$

← 用(b, w) 做 求得的 prediction.

Are We Done Yet?

goal: SVM **without dependence on $\tilde{d}$**

$$\min_{\boldsymbol{\alpha}} \quad \frac{1}{2}\boldsymbol{\alpha}^T \mathrm{Q_D} \boldsymbol{\alpha} - \mathbf{1}^T \boldsymbol{\alpha}$$

$$\text{subject to} \quad \mathbf{y}^T \boldsymbol{\alpha} = 0;$$

跟$\tilde{d}$有关

$$\alpha_n \geq 0, \text{for } n = 1, 2, \ldots, N$$

- N variables, $N + 1$ constraints: **no dependence on $\tilde{d}$?**
- $\boxed{q_{n,m} = y_n y_m \mathbf{z}_n^T \mathbf{z}_m}$: inner product in $\mathbb{R}^{\tilde{d}}$ —$O(\tilde{d})$ via naïve computation!

如何避开$\tilde{d}$的计算

no dependence only if
avoiding naïve computation **(next lecture :-))**

Questions?

Summary

1. Embedding Numerous Features: Kernel Models

Lecture 10: Support Vector Machine (1)

- Large-Margin Separating Hyperplane
  intuitively more robust against noise
- Standard Large-Margin Problem
  minimize 'length of w' at special separating scale
- Support Vector Machine
  'easy' via quadratic programming
- Motivation of Dual SVM
  want to remove dependence on $\tilde{d}$
- Lagrange Dual SVM
  KKT conditions link primal/dual
- Solving Dual SVM
  another QP, better solved with special solver
- Messages behind Dual SVM
  SVs represent fattest hyperplane