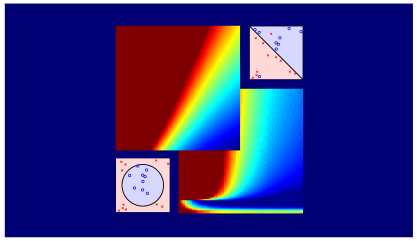
Machine Learning Techniques
(機器學習技巧)



Lecture 1: Large-Margin Linear Classification

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)

Agenda

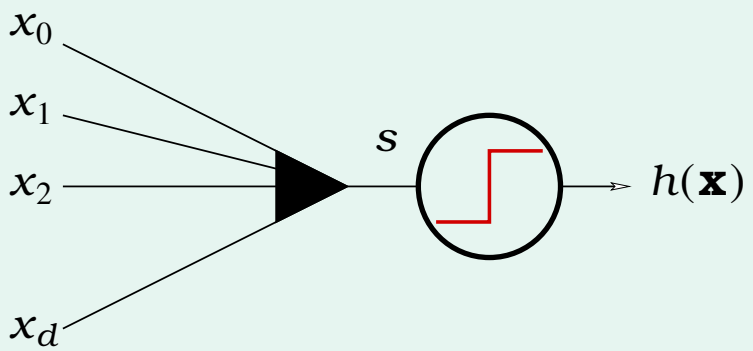
Lecture 1: Large-Margin Linear Classification

- Large-Margin Separating Hyperplane
- Standard Large-Margin Problem
- Support Vector Machine
- Reasons behind Large-Margin Hyperplane

Linear Classification Revisited

PLA/pocket

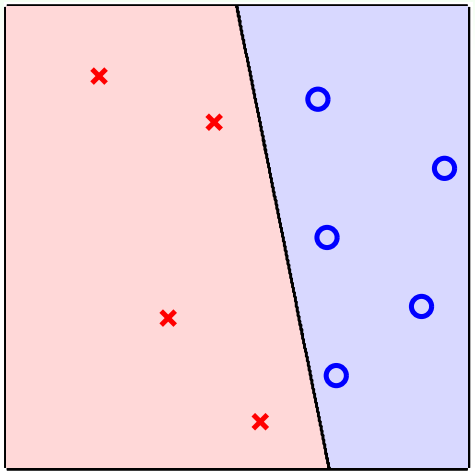
$$h(\mathbf{x}) = \text{sign}(s)$$



plausible err = 0/1

(small flipping noise)

minimize **specially**

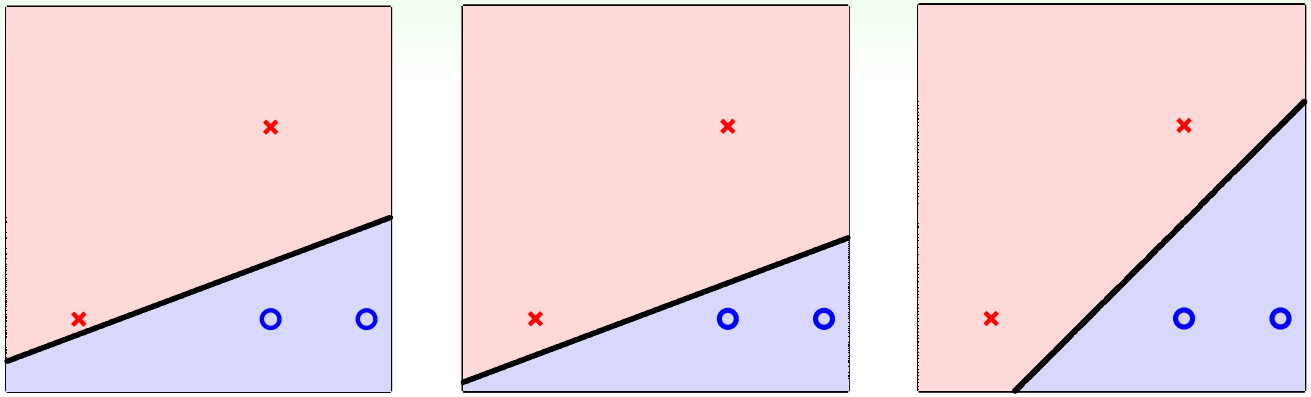


(linear separable)

linear (hyperplane) classifiers:

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

Which Line Is Best?

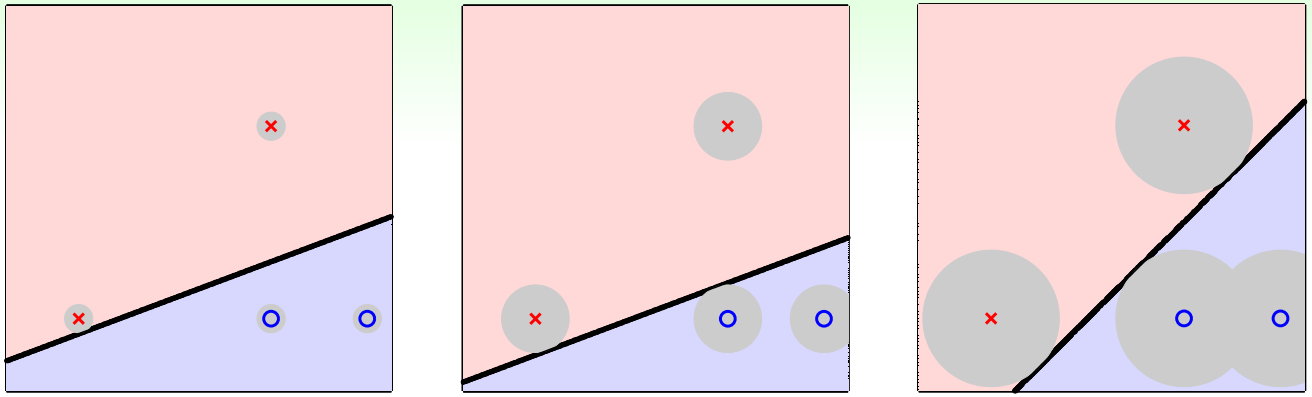


- PLA? depending on randomness
- VC bound? whichever you like!

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

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

Why Rightmost Hyperplane?

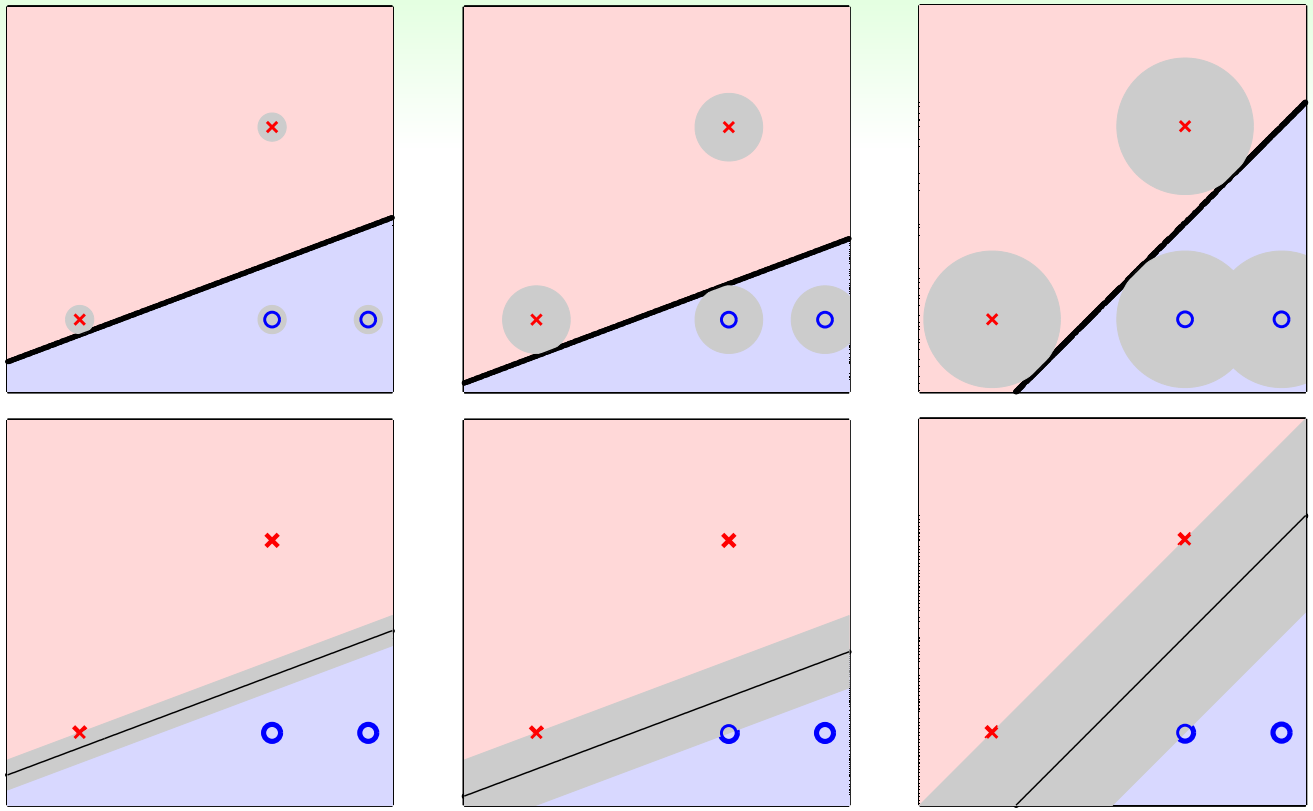


informal argument

- if (Gaussian-like) noise on future $\mathbf{x} \approx \mathbf{x}_n$:
  $\mathbf{x}_n$ further from hyperplane
  $\Longleftrightarrow$ tolerate more noise
  $\Longleftrightarrow$ more robust to overfitting
- robustness of separating hyperplane
  $\Longleftrightarrow$ amount of noise tolerance
  $\Longleftrightarrow$ distance to closest $\mathbf{x}_n$

rightmost one: **more robust**

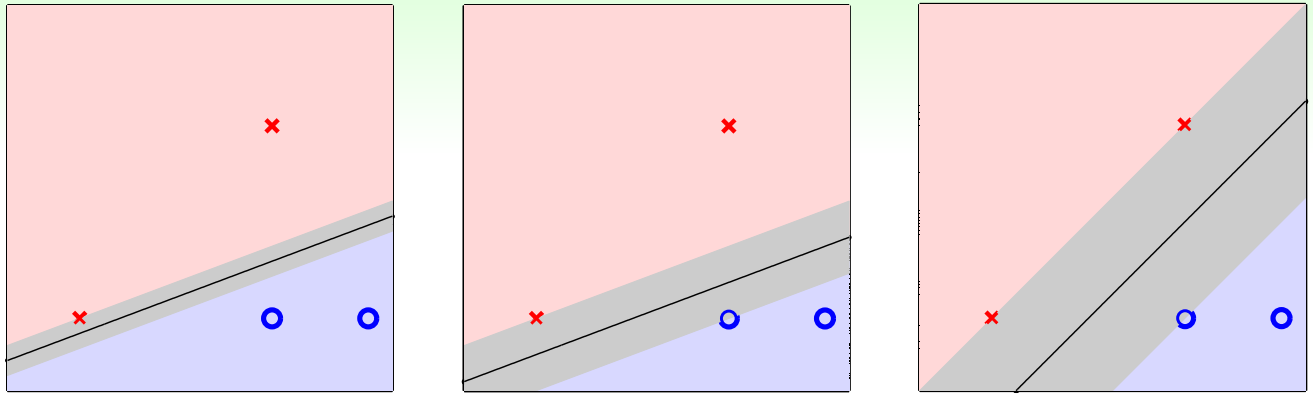
Fat Hyperplane



- 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

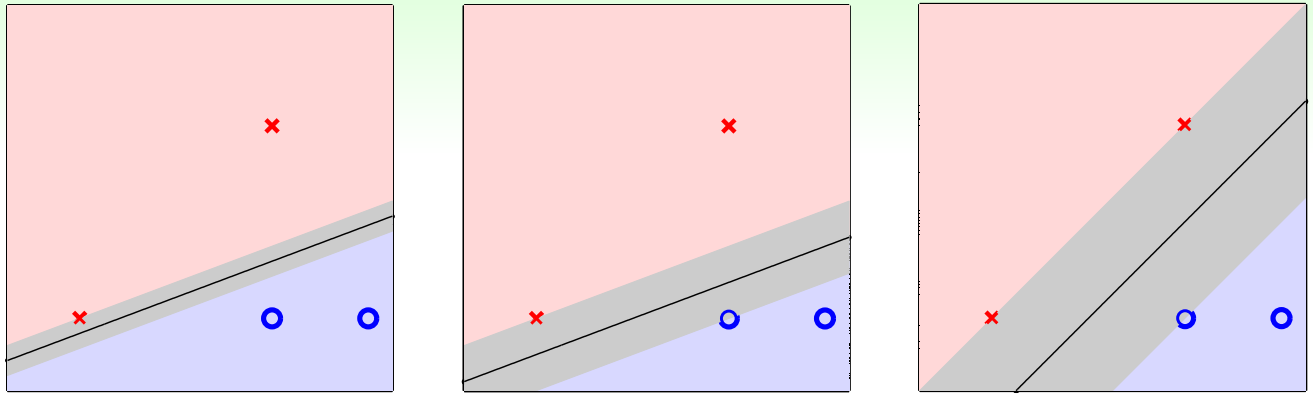


$$\begin{aligned} \max_{\mathbf{w}} \quad & \text{fatness}(\mathbf{w}) \\ \text{subject to} \quad & \mathbf{w} \text{ classifies every } (\mathbf{x}_n, y_n) \text{ correctly} \\ & \text{fatness}(\mathbf{w}) = \min_{n=1,\dots,N} \text{distance}(\mathbf{x}_n, \mathbf{w}) \end{aligned}$$

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

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

Large-Margin Separating Hyperplane



$$\begin{aligned} \max_{\mathbf{w}} \quad & \textbf{margin}(\mathbf{w}) \\ \text{subject to} \quad & \text{every } y_n \mathbf{w}^T \mathbf{x}_n > 0 \\ & \textbf{margin}(\mathbf{w}) = \min_{n=1,\dots,N} \text{distance}(\mathbf{x}_n, \mathbf{w}) \end{aligned}$$

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

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

Fun Time

Distance to Hyperplane: Preliminary

$$\begin{aligned} \max_{\mathbf{w}} \quad & \text{margin}(\mathbf{w}) \\ \text{subject to} \quad & \text{every } y_n \mathbf{w}^T \mathbf{x}_n > 0 \\ & \text{margin}(\mathbf{w}) = \min_{n=1,\dots,N} \text{distance}(\mathbf{x}_n, \mathbf{w}) \end{aligned}$$

'shorten' $\mathbf{x}$ and $\mathbf{w}$

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

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

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

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

Distance to Hyperplane

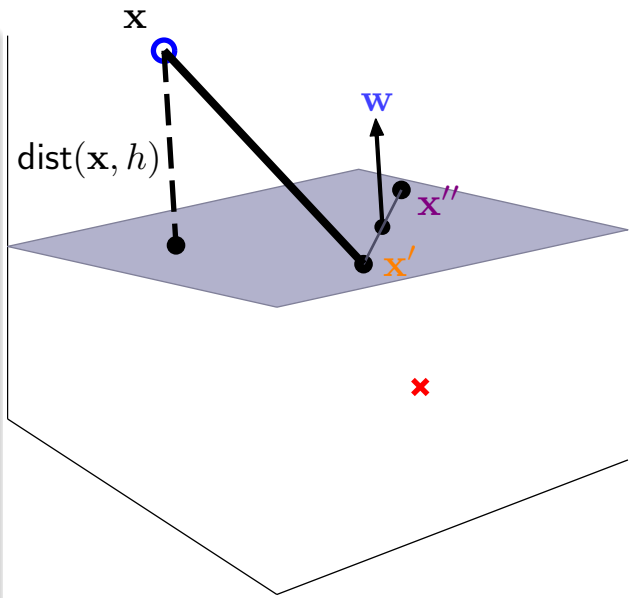
want: distance($\mathbf{x}, \mathbf{w}, b$), with hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$

consider $\mathbf{x}'$ on hyperplane

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

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

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



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

Distance to **Separating** Hyperplane

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

- **separating** hyperplane: for every n

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

- distance to **separating** hyperplane:

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

$$\begin{aligned} \max_{b,\mathbf{w}} \quad & \text{margin}(\mathbf{w}, b) \\ \text{subject to} \quad & \text{every } y_n(\mathbf{w}^T \mathbf{x}_n + b) > 0 \\ & \text{margin}(\mathbf{w}, b) = \min_{n=1,\dots,N} \frac{1}{\|\mathbf{w}\|} y_n(\mathbf{w}^T \mathbf{x}_n + b) \end{aligned}$$

Margin of **Special** Separating Hyperplane

$$\begin{aligned} \max_{b,\mathbf{w}} \quad & \text{margin}(\mathbf{w}, b) \\ \text{subject to} \quad & \text{every } y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0 \\ & \text{margin}(\mathbf{w}, b) = \min_{n=1,\dots,N} \frac{1}{\|\mathbf{w}\|} y_n(\mathbf{w}^T\mathbf{x}_n + b) \end{aligned}$$

- $(\mathbf{w}, b)$ and $(1126\mathbf{w}, 1126b)$: same hyperplane, same margin
- **special** scaling: only consider separating $(\mathbf{w}, b)$ such that

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

$$\begin{aligned} \max_{b,\mathbf{w}} \quad & \frac{1}{\|\mathbf{w}\|} \\ \text{subject to} \quad & \text{every } y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0 \\ & \min_{n=1,\dots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1 \end{aligned}$$

Standard Large-Margin Hyperplane Problem

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

final changes:

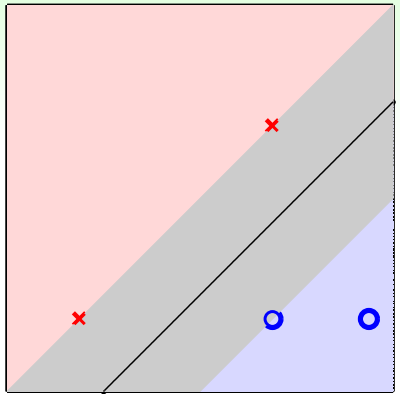
- max $\Longrightarrow$ min, remove $\sqrt{\ }$, add $\frac{1}{2}$
- $\min(\dots) = 1 \Longrightarrow (\dots) \geq 1$
  —min $\frac{1}{2}\mathbf{w}^T\mathbf{w}$ means not all $(\dots) > 1$

$$\min_{b,\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{subject to} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1$$

Fun Time

Solving a Particular Standard Problem

$$\begin{aligned} \min_{b,\mathbf{w}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 \end{aligned}$$



$$\mathrm{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix} \qquad \begin{aligned} -b &\geq 1 \quad (i) \\ -2w_1 - 2w_2 - b &\geq 1 \quad (ii) \\ 2w_1 + b &\geq 1 \quad (iii) \\ 3w_1 + b &\geq 1 \quad (iv) \end{aligned}$$

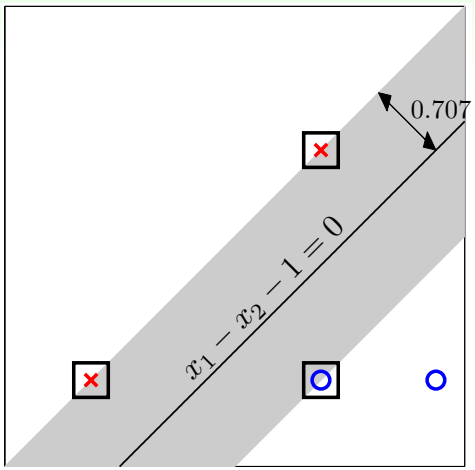
- $\left.\begin{array}{lll} (i) & \& & (iii) \\ (ii) & \& & (iii) \end{array}\right\} \begin{array}{c} \Longrightarrow \\ \Longrightarrow \end{array} \begin{array}{l} w_1 \geq +1 \\ w_2 \leq -1 \end{array} \Longrightarrow \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \geq 1$
- $(w_1 = 1, w_2 = -1, b = -1)$ at **lower bound** and satisfies $(i) - (iv)$

$$g_{\mathrm{SVM}}(\mathbf{x}) = \mathrm{sign}(x_1 - x_2 - 1)\text{: SVM? :-)}$$

Support Vector Machine (SVM)

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

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



- examples on boundary: 'locates' fattest hyperplane
  other examples: **not needed**
- call boundary examples **support vector** (candidates)

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

Solving General SVM

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

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

- **not easy manually, of course :-)**
- gradient decent? **not easy with constraints**
- luckily:
  - (convex) quadratic objective functions of $(b, \mathbf{w})$
  - linear constraints of $(b, \mathbf{w})$

—**quadratic programming**

quadratic programming (QP):
'easy' optimization problem

Quadratic Programming

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

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

$$\text{subject to} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \quad \text{for } n = 1, 2, \ldots, N$$

optimal $\mathbf{u} \leftarrow \text{QP}(\mathrm{A}, \mathbf{c}, \mathrm{P}, \mathbf{r})$

$$\min_{\mathbf{u}} \quad \frac{1}{2}\mathbf{u}^T\mathrm{A}\mathbf{u} + \mathbf{c}^T\mathbf{u}$$

$$\text{subject to} \quad \mathbf{p}_m^T\mathbf{u} \geq r_m, \quad \text{for } m = 1, 2, \ldots, M$$

objective function: $\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}; \mathrm{A} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathrm{I}_d \end{bmatrix}; \mathbf{c} = \mathbf{0}_{d+1}$

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

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

SVM with QP Solver

Linear Hard-Margin SVM Algorithm

1. $A = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & I_d \end{bmatrix}$; $\mathbf{c} = \mathbf{0}_{d+1}$; $\mathbf{p}_n^T = y_n \begin{bmatrix} 1 & \mathbf{x}_n^T \end{bmatrix}$; $r_n = 1$
2. $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \text{QP}(A, \mathbf{c}, P, \mathbf{r})$
3. return b & $\mathbf{w}$ as g_{SVM}

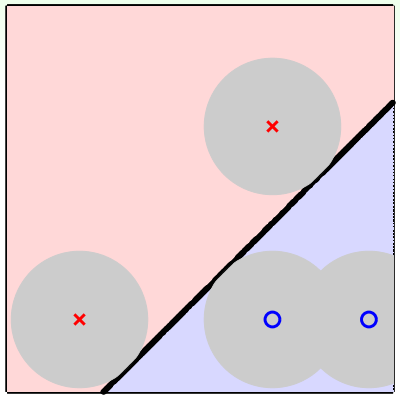
- **hard-margin**: nothing violate 'fat boundary'
- **linear**: $\mathbf{x}_n$

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

Fun Time

Why Large-Margin Hyperplane?

$$\begin{aligned} \min_{b,\mathbf{w}} \quad & \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \\ \text{subject to} \quad & y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1 \end{aligned}$$



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

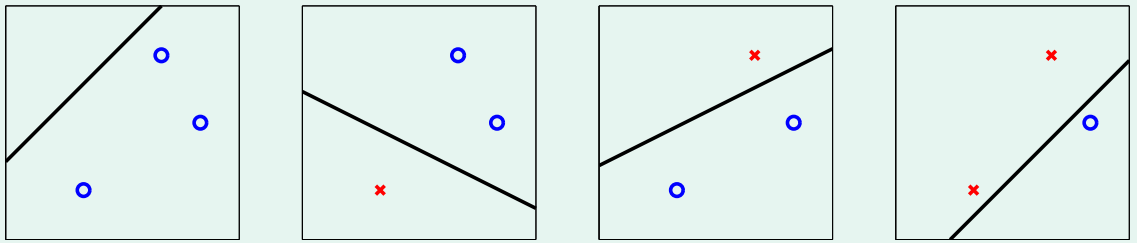
SVM (large-margin hyperplane):
'weight-decay regularization' within $E_{\text{in}} = 0$

Large-Margin Restricts Dichotomies

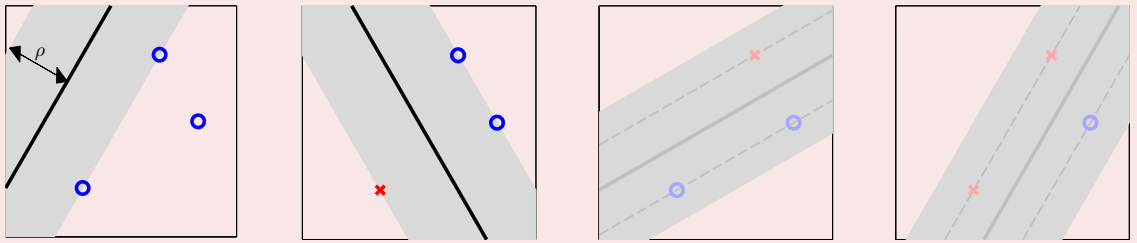
consider 'large-margin algorithm' $\mathcal{A}_\rho$:

either **returns g with margin$(g) \geq \rho$ (if exists)**, or 0 otherwise

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



$\mathcal{A}_{1.126}$: more strict than SVM $\Longrightarrow$ no-shatter some 3 inputs



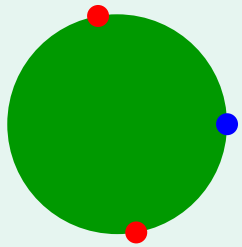
fewer dichotomies $\Longrightarrow$ smaller 'VC dim.' $\Longrightarrow$ **better generalization**

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]

$d_{\text{VC}}(\mathcal{A}_\rho)$ when $\mathcal{X}$ = unit circle in $\mathbb{R}^2$

- $\rho = 0$: just perceptrons ($d_{\text{VC}} = 3$)
- $\rho > \frac{\sqrt{3}}{2}$: no shatter any 3 inputs ($d_{\text{VC}} < 3$)
  —some inputs must be of **distance** $\leq \sqrt{3}$



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

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

Benefits of Large-Margin Hyperplanes

	large-margin hyperplanes	hyperplanes	hyperplanes + higher-order transforms
#	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 + higher-order transforms
#	**not many**
boundary	**sophisticated**

Fun Time

Summary

Lecture 1: Large-Margin Linear Classification

- Large-Margin Separating Hyperplane
  intuitively more robust
- Standard Large-Margin Problem
  min. normal vector while separating with scale
- Support Vector Machine
  easy via quadratic programming
- Reasons behind Large-Margin Hyperplane
  fewer dichotomies and better generalization