Three Problems

Problem 1: Evaluation

- What does F(x,y) look like?

Problem 2: Inference

- How to solve the “arg max

$y = \arg m$

Problem 3: Training

- Given training data, how t

Have you heard the three problems elsewhere?

Hidden Markov Model

- **Three Basic Problems for HMMs**

Given an observation sequence $\overline{O}=(o_1,o_2,\ldots,o_T)$, and an HMM $\lambda=(A,B,\pi)$

– Problem *1* :

How to *efficiently* compute $P(\overline{O}|\lambda)$?

⇨ *Evaluation problem*

– Problem *2* :

How to choose an optimal state sequence $\mathbf{q}=(q_1,q_2,\ldots\ldots, q_T)$?

⇨ *Decoding Problem*

– Problem *3* :

Given some observations $\overline{O}$ for the HMM λ , how to adjust the model parameter $\lambda=(A,B,\pi)$ to maximize $P(\overline{O}|\lambda)$?

⇨ *Learning /Training Problem*

HMM其實就是structure learning的一個特例

From 數位語音處理

Link to DNN?

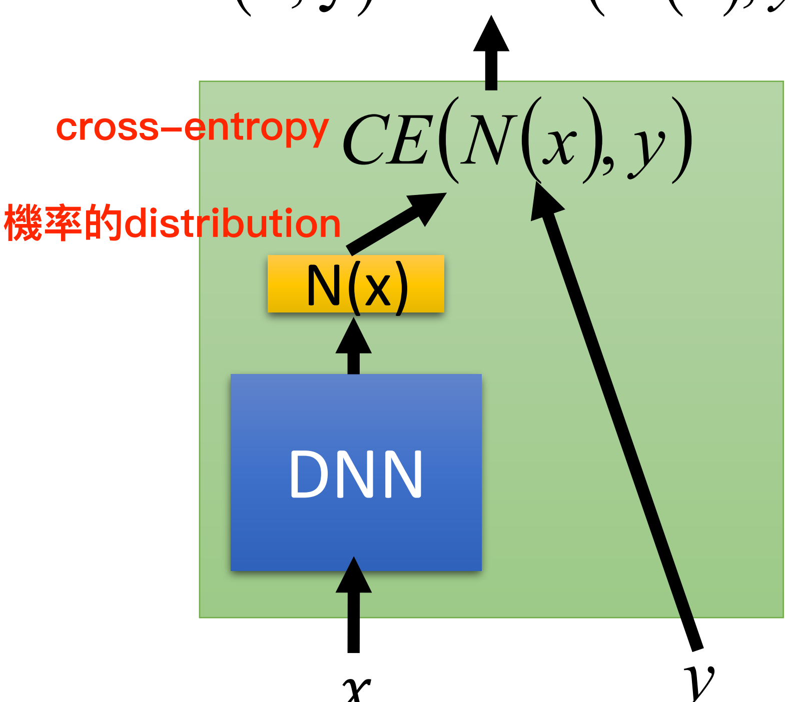
The same as what we have learned.

Training

$$\mathrm{F}: X \times Y \rightarrow \mathrm{R}$$

CE要越小越好

$$\mathrm{F}(\mathrm{x},\mathrm{y}) = -CE(\mathrm{N}(x), y)$$



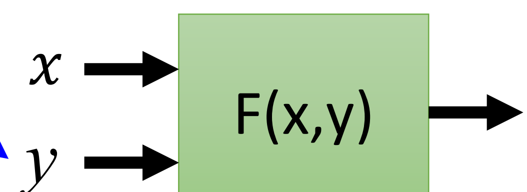
Inference

$$\tilde{y} = \arg\max_{y \in Y} F(x, y)$$

In handwriting digit classification, there are only 10 possible y.

y = [1 0 0 0 ......]
y = [0 1 0 0 ......]
y = [0 0 1 0 ......]
⋮

Find max



Introduction of Structured Learning

Linear Model

Structured Linear Model

做出一個限制：假設evaluation function是linear

Problem 1: Evaluation

- What does F(x,y) look like? ➡ in a specific form

Problem 2: Inference

- How to solve the "arg max" problem

$$y = \arg\max_{y \in Y} F(x, y)$$

~~**Problem 3: Training**~~

- ~~Given training data, how to find F(x,y)~~

Structured Linear Model: Problem 1

限制假設為linear

- Evaluation: What does F(x,y) look like?

Characteristics

x y

$$\phi_1(x,y)$$
$$\phi_2(x,y)$$
$$\phi_3(x,y)$$
$$\vdots$$

抽取不同的feature做比較

$$\mathrm{F}(x,y) = \boxed{w_1} \cdot \phi_1(x,y) + \boxed{w_2} \cdot \phi_2(x,y) + \boxed{w_3} \cdot \phi_3(x,y) \ldots$$

Learning from data

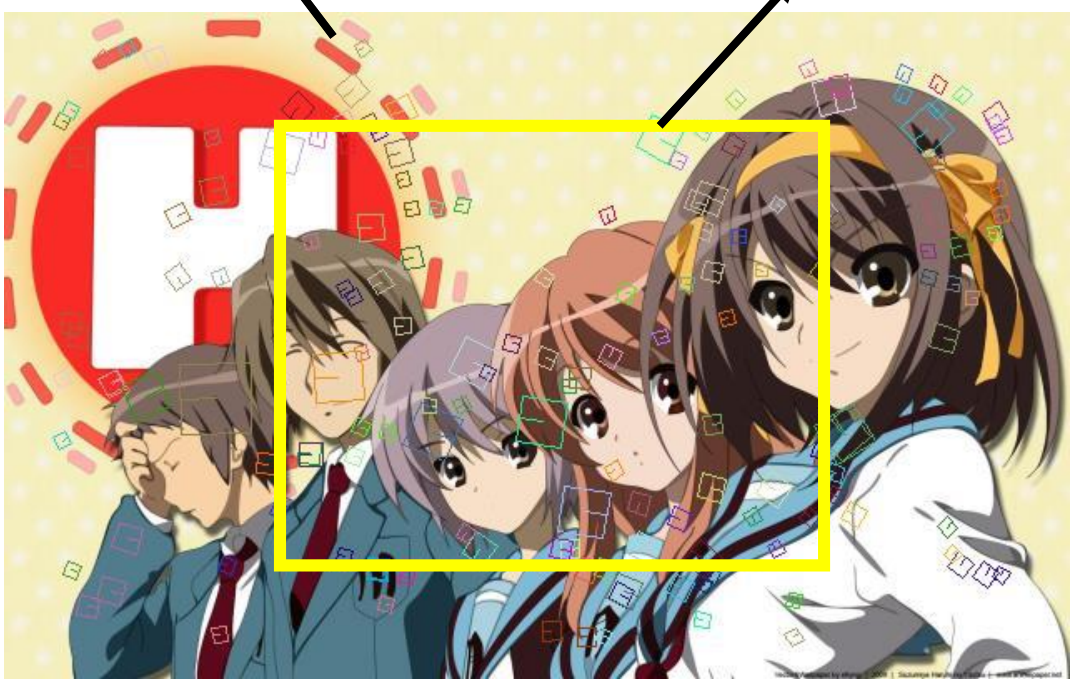
$$\mathrm{F}(x,y) = \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \end{bmatrix}}_{w} \cdot \underbrace{\begin{bmatrix} \phi_1(x,y) \\ \phi_2(x,y) \\ \phi_3(x,y) \\ \vdots \end{bmatrix}}_{\phi(x,y)}$$

$$\mathrm{F}(x,y) = w \cdot \phi(x,y)$$

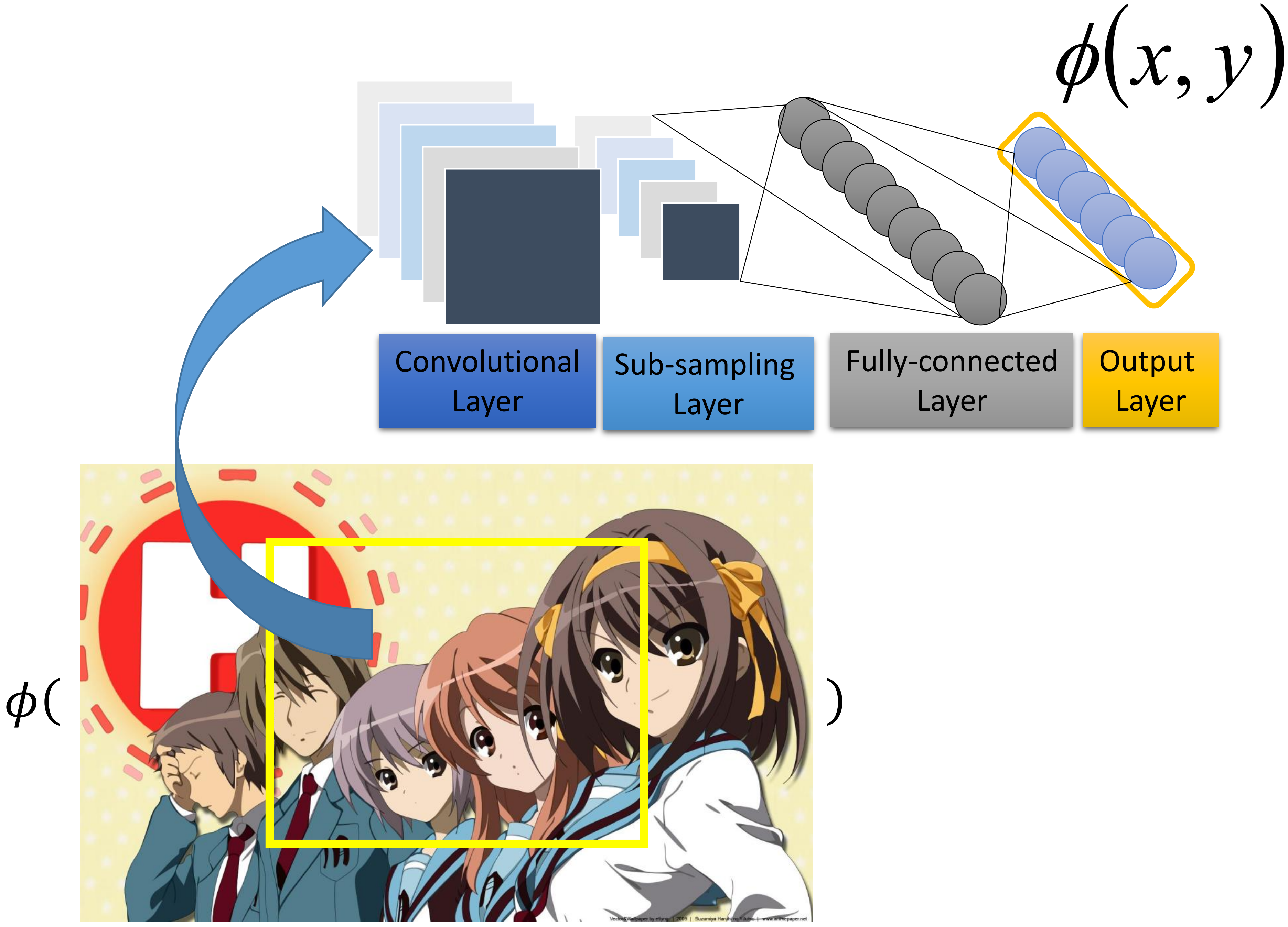
Structured Linear Model: Problem 1

- Evaluation: What does F(x,y) look like?
- Example: ***Object Detection***

x $\quad$ y

$\phi($  $) =$

$$\begin{bmatrix} \text{percentage of color red in box } y \\ \text{percentage of color green in box } y \\ \text{percentage of color blue in box } y \\ \text{percentage of color red out of box } y \\ \ldots\ldots \\ \text{area of box } y \\ \text{number of specific patterns in box } y \\ \ldots\ldots \end{bmatrix}$$


$\phi(x, y)$
Convolutional Layer
Sub-sampling Layer
Fully-connected Layer
Output Layer
$\phi($ $)$

Structured Linear Model: Problem 2

- **Inference**: How to solve the "arg max" problem

$$y = \arg\max_{y \in Y} F(x, y)$$

$$F(x, y) = w \cdot \phi(x, y) \Rightarrow y = \arg\max_{y \in Y} w \cdot \phi(x, y)$$

- Assume we have solved this question.

假設第一步是linear第二步也解出來則第三步是很簡單的

Structured Linear Model: Problem 3

- Training: Given training data, how to learn F(x,y)
  - F(x,y) = w·ϕ(x,y), so what we have to learn is w

Training data: $\left\{\left(x^1,\hat{y}^1\right),\left(x^2,\hat{y}^2\right),\ldots,\left(x^r,\hat{y}^r\right),\ldots\right\}$

We should find w such that

$\forall r$ (All training examples)

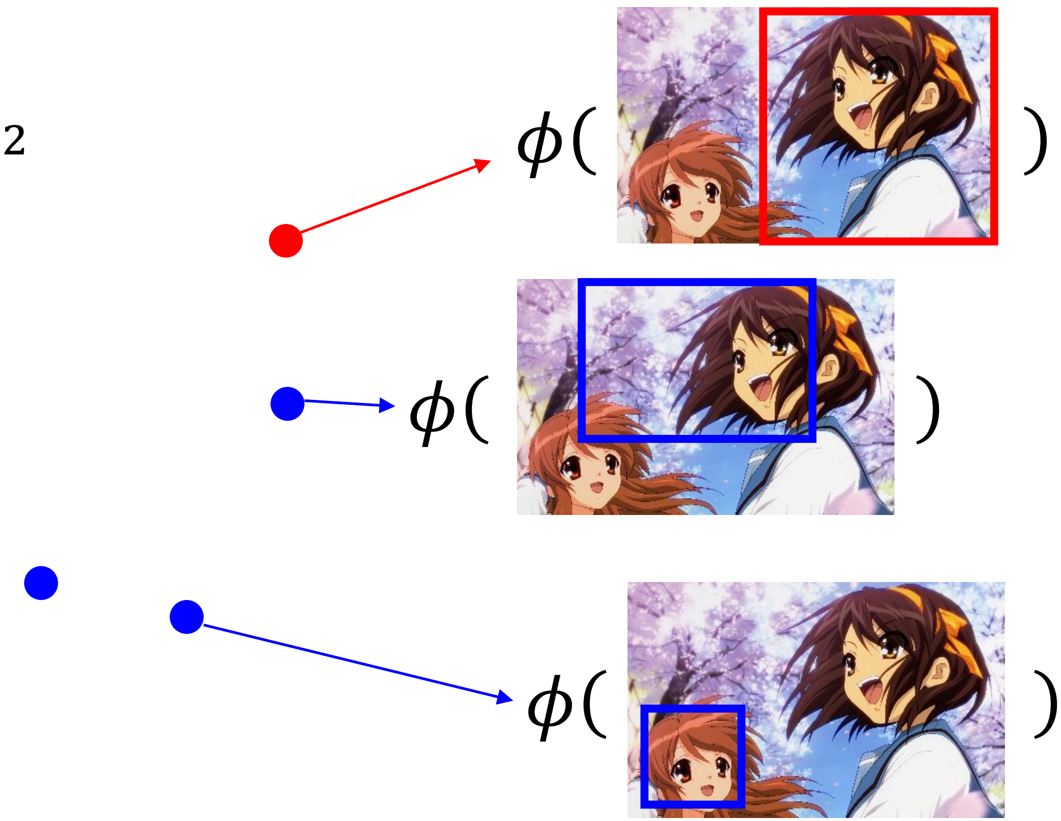
$\forall y \in Y - \{\hat{y}^r\}$ (All incorrect label for r-th example)

$$w \cdot \phi\left(x^r,\hat{y}^r\right) > w \cdot \phi\left(x^r, y\right)$$

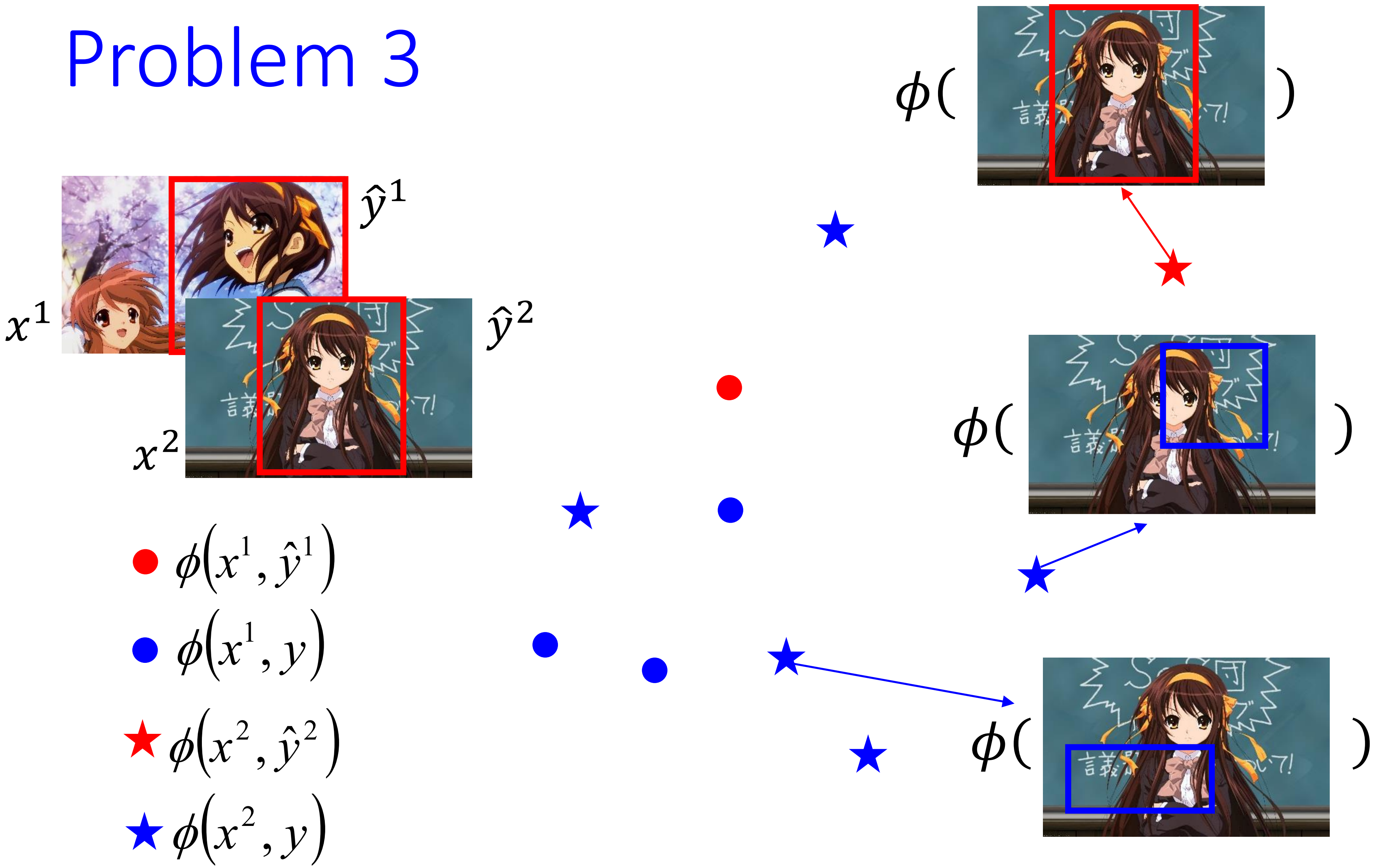
Structured Linear Model: Problem 3

正確的 ● $\phi\left(x^1, \hat{y}^1\right)$

錯誤的 ● $\phi\left(x^1, y\right)$



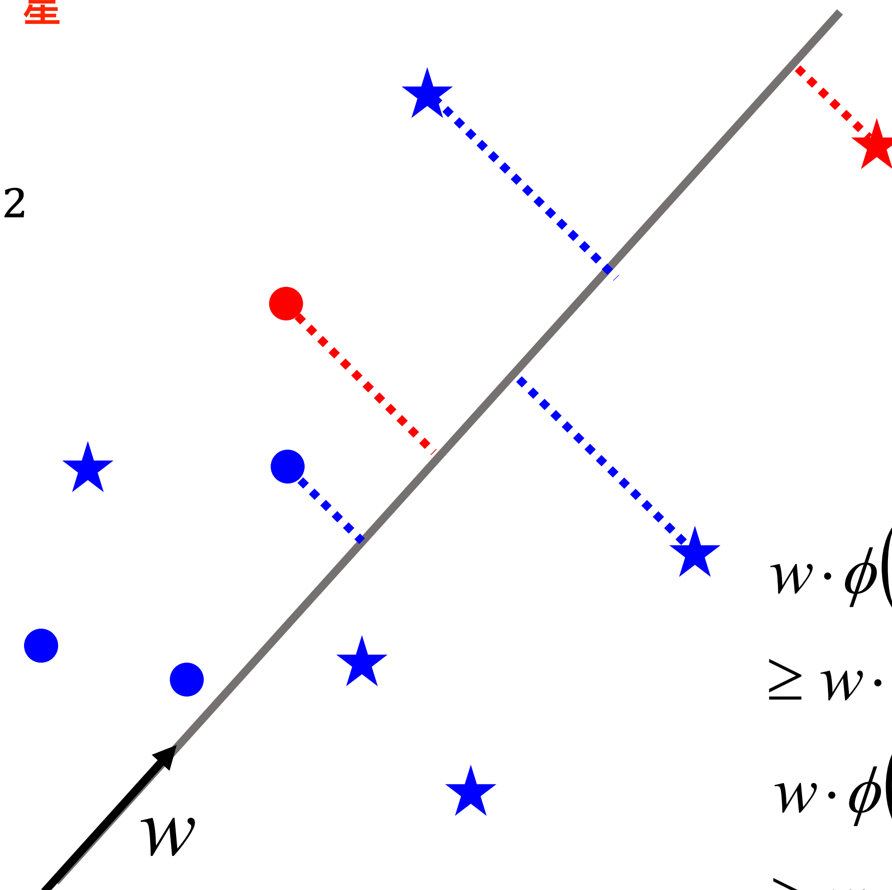
Structured Linear Model:
Problem 3



Structured Linear Model: Problem 3

紅色圈贏過所有藍色圈
紅色星贏過所有藍色星

圈圈跟星星間彼此不比較



$\color{red}\bullet$ $\phi(x^1, \hat{y}^1)$

$\color{blue}\bullet$ $\phi(x^1, y)$

$\color{red}\bigstar$ $\phi(x^2, \hat{y}^2)$

$\color{blue}\bigstar$ $\phi(x^2, y)$

$$w \cdot \phi(x^1, \hat{y}^1) \geq w \cdot \phi(x^1, y)$$

$$w \cdot \phi(x^2, \hat{y}^2) \geq w \cdot \phi(x^2, y)$$

Solution of Problem 3

Difficult?

Not as difficult as expected

Algorithm

Will it terminate?

- **Input**: training data set $\{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \ldots, (x^r, \hat{y}^r), \ldots\}$
- **Output**: weight vector w
- **Algorithm**: Initialize w = 0
  - do
    - For each pair of training example $(x^r, \hat{y}^r)$
      - Find the label $\tilde{y}^r$ maximizing $w \cdot \phi(x^r, y)$

解這個optimization的problem $\tilde{y}^r = \arg\max_{y \in Y} w \cdot \phi(x^r, y)$ (question 2)

      - If $\tilde{y}^r \neq \hat{y}^r$, update w

如果找出來不是正確答案，則做更新
如果找出來是正確的則這個式子會不增不減

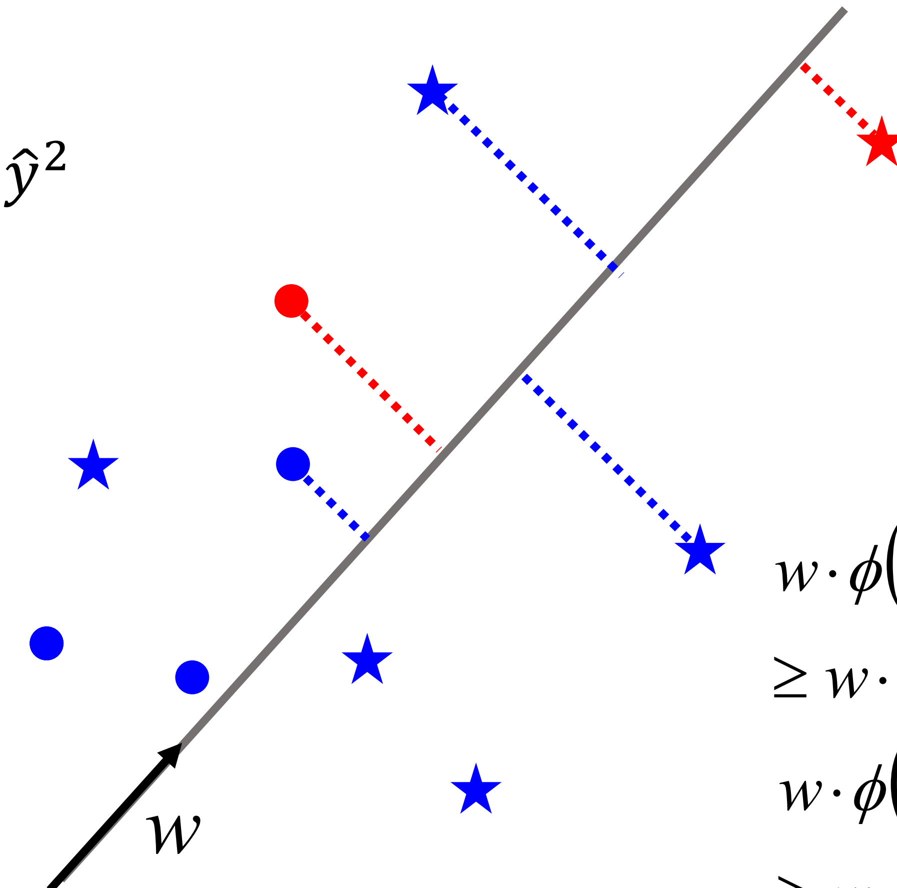
$$w \rightarrow w + \phi(x^r, \hat{y}^r) - \phi(x^r, \tilde{y}^r)$$

  - until w is not updated → We are done!

對所有train example，每個pair都是正確的話則結束

Algorithm - Example

- ● $\phi(x^1, \hat{y}^1)$
- ● $\phi(x^1, y)$
- ★ $\phi(x^2, \hat{y}^2)$
- ★ $\phi(x^2, y)$



$$w \cdot \phi(x^1, \hat{y}^1) \geq w \cdot \phi(x^1, y)$$

$$w \cdot \phi(x^2, \hat{y}^2) \geq w \cdot \phi(x^2, y)$$

● $\phi(x^1, \hat{y}^1)$ (red)

● $\phi(x^1, y)$ (blue)

★ $\phi(x^2, \hat{y}^2)$ (red)

★ $\phi(x^2, y)$ (blue)

Algorithm - Example

Initialize w = 0

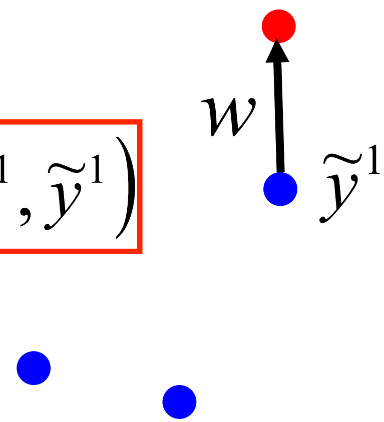
pick $(x^1, \hat{y}^1)$ 隨機選一個 example pair

$$\tilde{y}^1 = \arg\max_{y \in Y} w \cdot \phi(x^1, y)$$

If $\tilde{y}^1 \neq \hat{y}^1$, update w

$$w \rightarrow w + \boxed{\phi(x^1, \hat{y}^1) - \phi(x^1, \tilde{y}^1)}$$

vector



Because w=0 at this time, $\phi(x^1, y)$ always 0

因為w初始為零，因此隨機選一個y

➡ Random pick one point as $\tilde{y}^r$

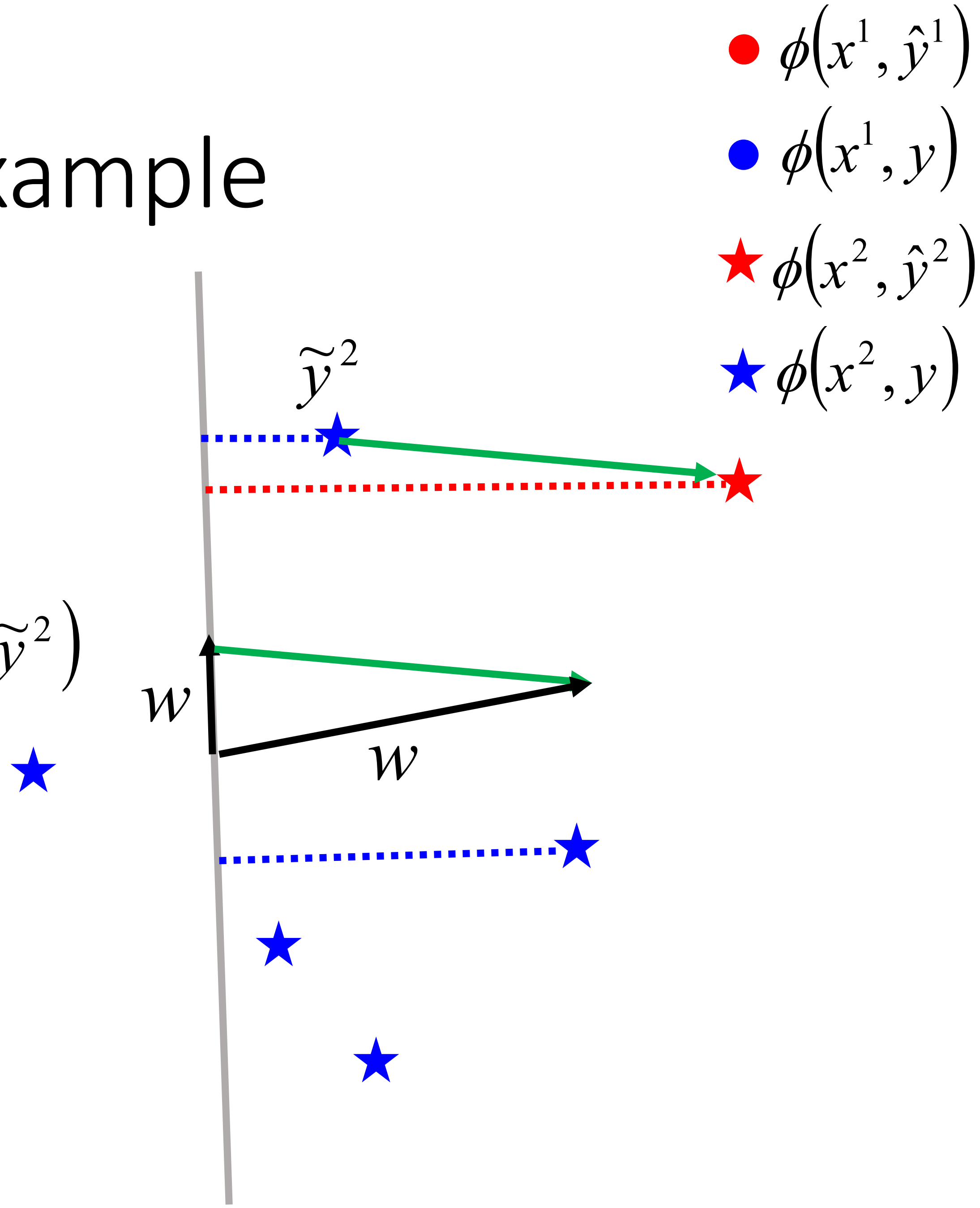
Algorithm - Example

pick $(x^2, \hat{y}^2)$

$$\tilde{y}^2 = \arg\max_{y \in Y} w \cdot \phi(x^2, y)$$

If $\tilde{y}^2 \neq \hat{y}^2$, update w

$$w \rightarrow w + \phi(x^2, \hat{y}^2) - \phi(x^2, \tilde{y}^2)$$



Algorithm - Example

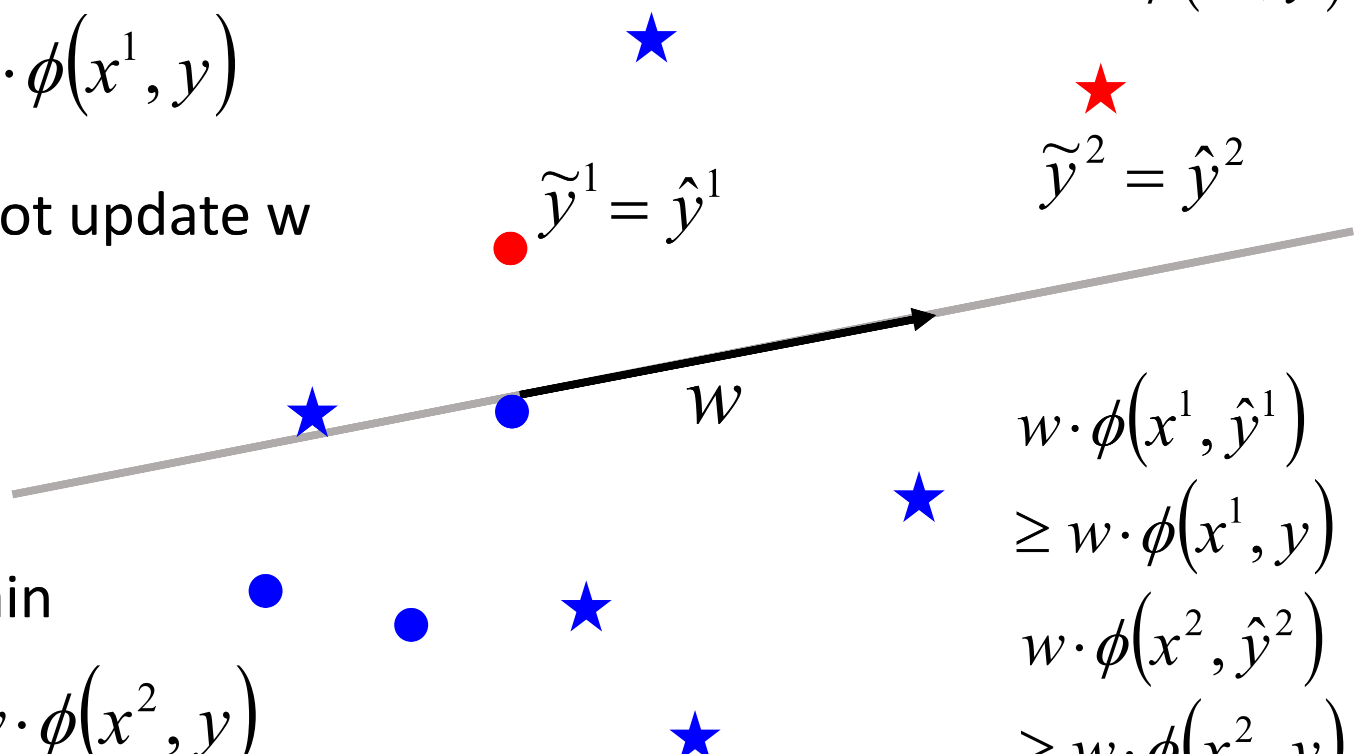
- ● $\phi(x^1, \hat{y}^1)$
- ● $\phi(x^1, y)$
- ★ $\phi(x^2, \hat{y}^2)$
- ★ $\phi(x^2, y)$

藍色的星星跟圈圈是有無限多個的，但是只要能解arg max的problem，則依然可以輕易解

pick $(x^1, \hat{y}^1)$ again

$$\tilde{y}^1 = \arg\max_{y \in Y} w \cdot \phi(x^1, y)$$

$\tilde{y}^1 = \hat{y}^1$ ➡ do not update w



$$w \cdot \phi(x^1, \hat{y}^1) \geq w \cdot \phi(x^1, y)$$

$$w \cdot \phi(x^2, \hat{y}^2) \geq w \cdot \phi(x^2, y)$$

pick $(x^2, \hat{y}^2)$ again

$$\tilde{y}^2 = \arg\max_{y \in Y} w \cdot \phi(x^2, y)$$

$\tilde{y}^2 = \hat{y}^2$ ➡ do not update w

So we are done

證明會收斂

Assumption: Separable

- There exists a weight vector $\hat{w}$ $\qquad \|\hat{w}\| = 1$

$\forall r$ (All training examples)

$\forall y \in Y - \{\hat{y}^r\}$ (All incorrect label for an example)

想要達成的目標

$$\hat{w} \cdot \phi(x^r, \hat{y}^r) \geq \hat{w} \cdot \phi(x^r, y)$$ (The target exists)

$$\hat{w} \cdot \phi(x^r, \hat{y}^r) \geq \hat{w} \cdot \phi(x^r, y) + \delta$$

Assumption: Separable

$$\hat{w} \cdot \phi\left(x^r, \hat{y}^r\right) \geq \hat{w} \cdot \phi\left(x^r, y\right) + \delta$$

- $\phi\left(x^1, \hat{y}^1\right)$
- $\phi\left(x^1, y\right)$
- $\phi\left(x^2, \hat{y}^2\right)$
- $\phi\left(x^2, y\right)$

……