Structured Linear Model

Hung-yi Lee

Structured Linear Model

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

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

Characteristics

$$\text{x} \quad \text{y} \rightarrow \begin{cases} \phi_1(x,y) \\ \phi_2(x,y) \\ \phi_3(x,y) \\ \vdots \end{cases}$$

$$\begin{aligned} \mathrm{F}(x,y) = & \; w_1 \cdot \phi_1(x,y) \\ & + w_2 \cdot \phi_2(x,y) \\ & + w_3 \cdot \phi_3(x,y) \ldots \end{aligned}$$

Learning from data

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

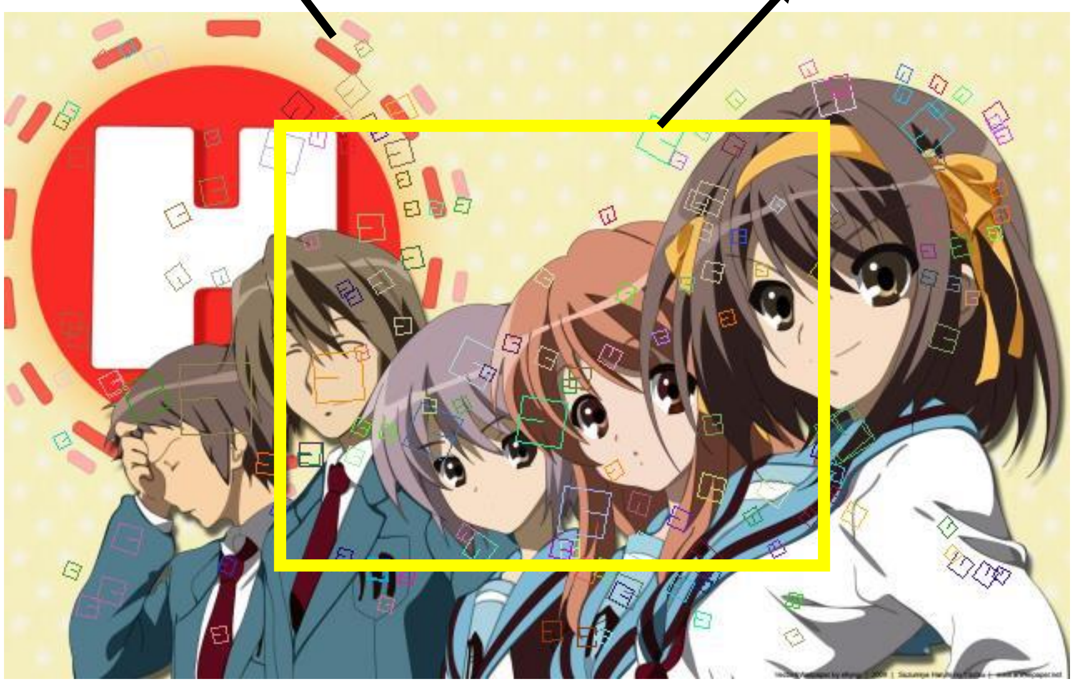
$$\mathrm{F}(x,y) = w \cdot \phi(x,y)$$

Structured Linear Model: Problem 1

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

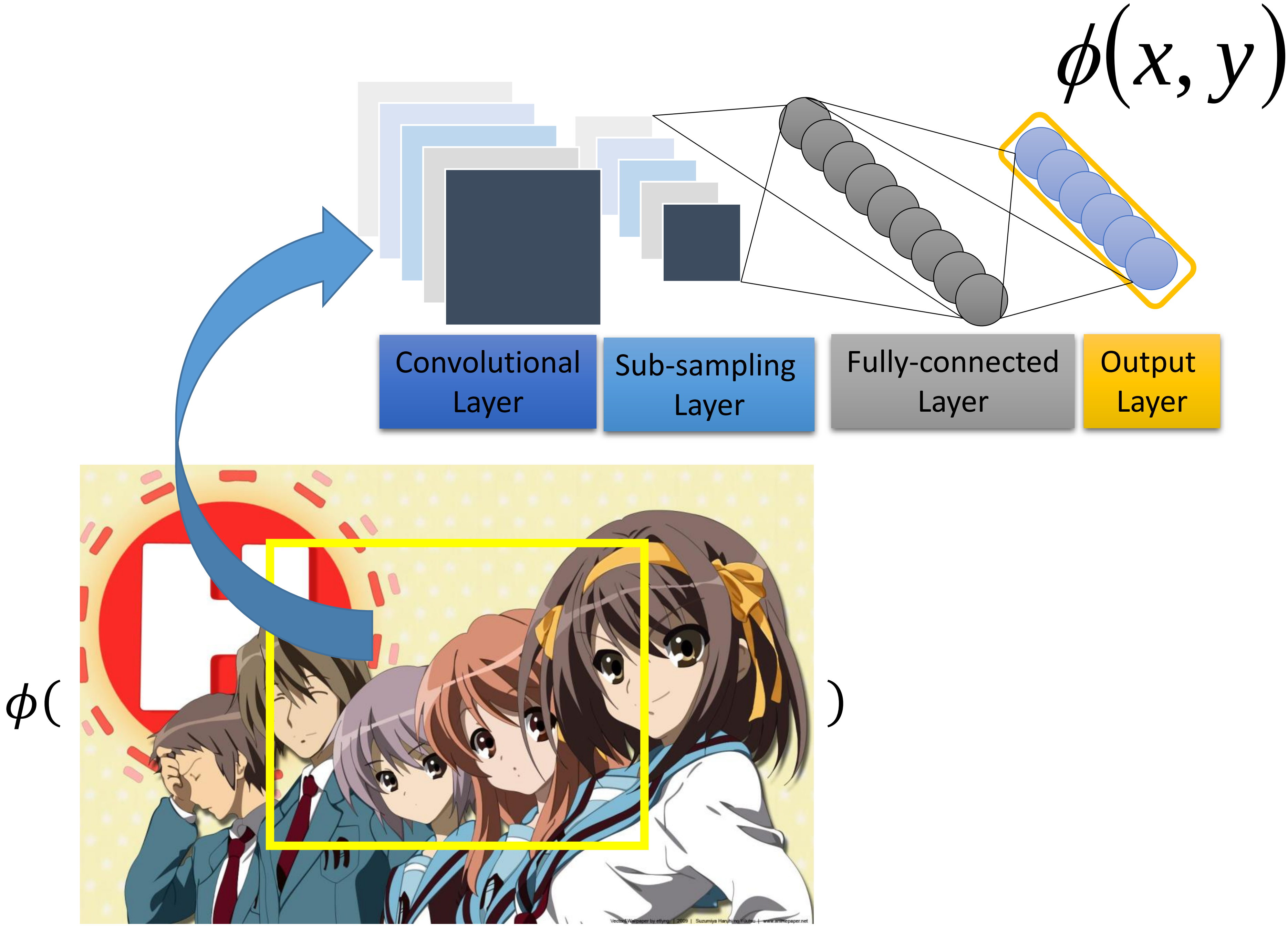
x y

$\phi($



$) =$

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


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

Structured Linear Model: Problem 1

- Evaluation: What does F(x,y) look like?
- Example: ***<u>Summarization</u>***

x **y**

(a long document)

(Short paragraph)

$\phi_1(x, y)$ → Whether the sentence containing the word "*important*" is in y

$\phi_2(x, y)$ → Whether the sentence containing the word "*definition*" is in y

$\phi_3(x, y)$ → Length of y

$\phi_4(x, y)$ → How succinct is y?

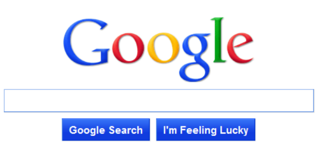
$\phi_5(x, y)$ → How representative of y?

$\vdots$

Structured Linear Model: Problem 1

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

x y



(Input keyword)

(Search Result)

$$\phi_1(x, y)$$

The degree of relevance with respect to x for the top 1 webpages in y.

$$\phi_2(x, y)$$

Is the top 1 webpage more relevant than the top 2 webpage?

$$\phi_3(x, y)$$

How much different information does y cover? (***Diversity***)

$$\vdots$$

Structured Linear Model: Problem 2

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

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

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

- Assume we have solved this question.

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

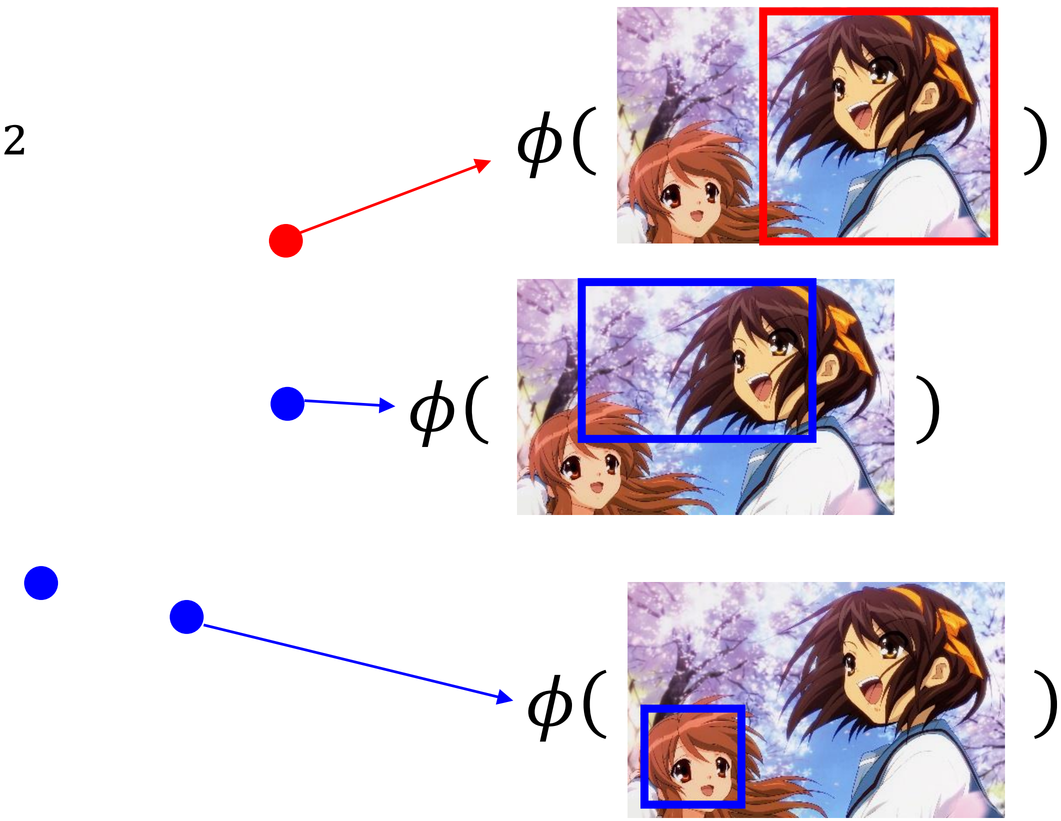
$\forall r$ (All training examples)

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

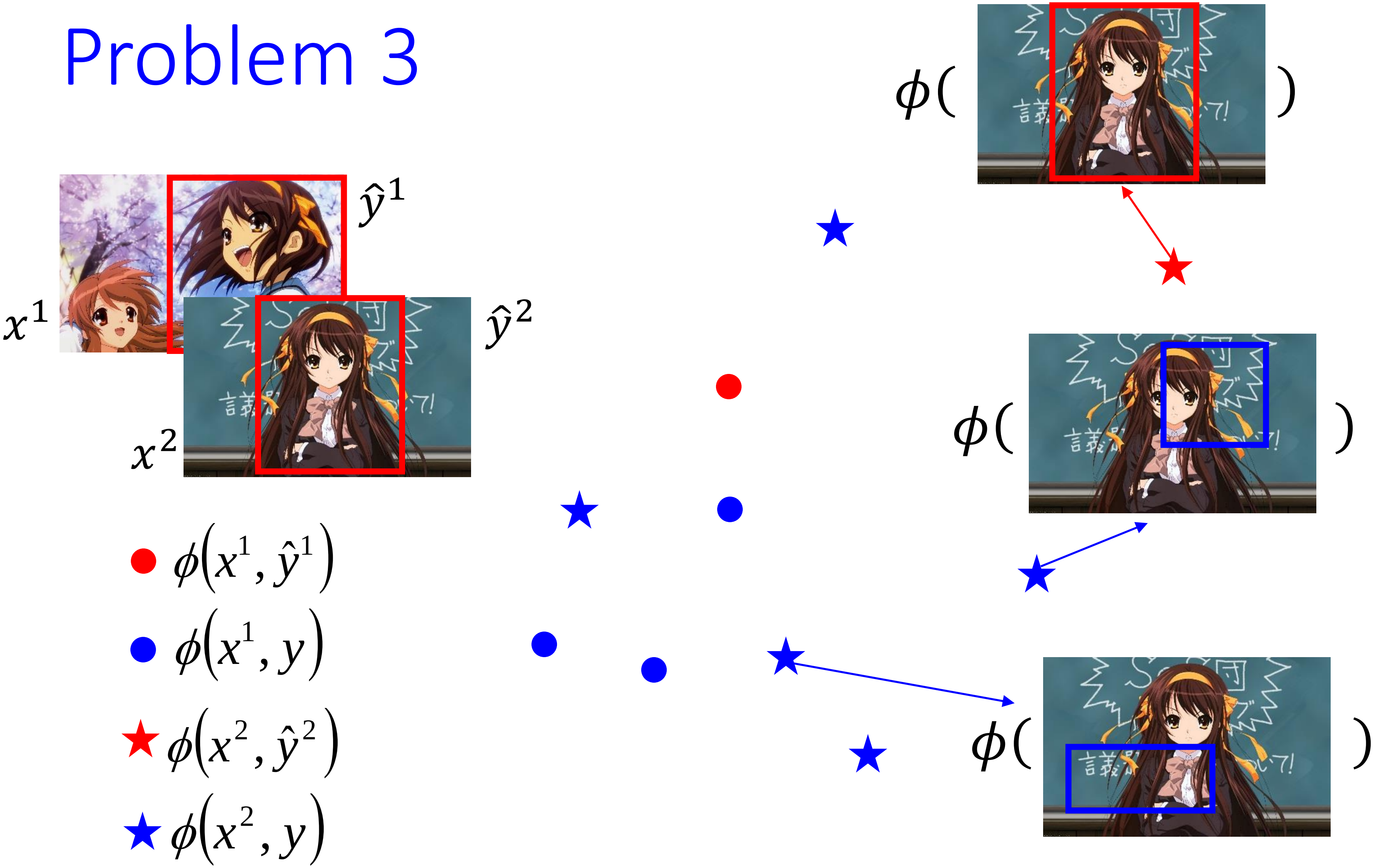
$$w \cdot \phi\left(x^r, \hat{y}^r\right) > w \cdot \phi\left(x^r, y\right)$$

Structured Linear Model: Problem 3

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



Structured Linear Model: Problem 3

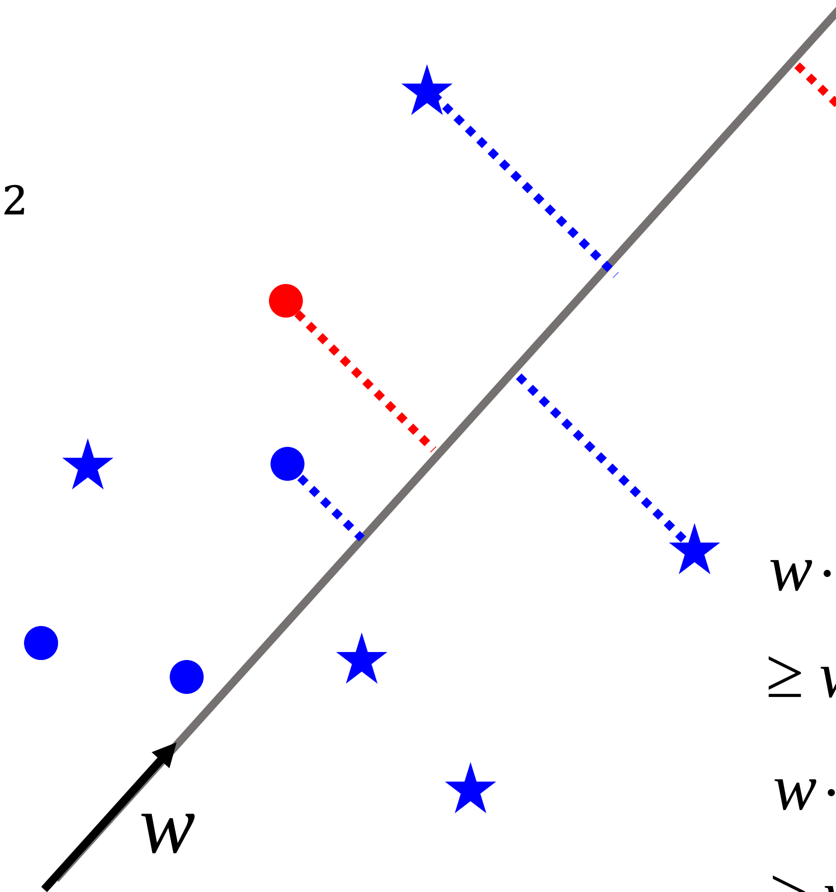


Structured Linear Model: Problem 3

x^1 $\hat{y}^1$

x^2 $\hat{y}^2$

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



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

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

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 $w \cdot \phi(x^r, y)$

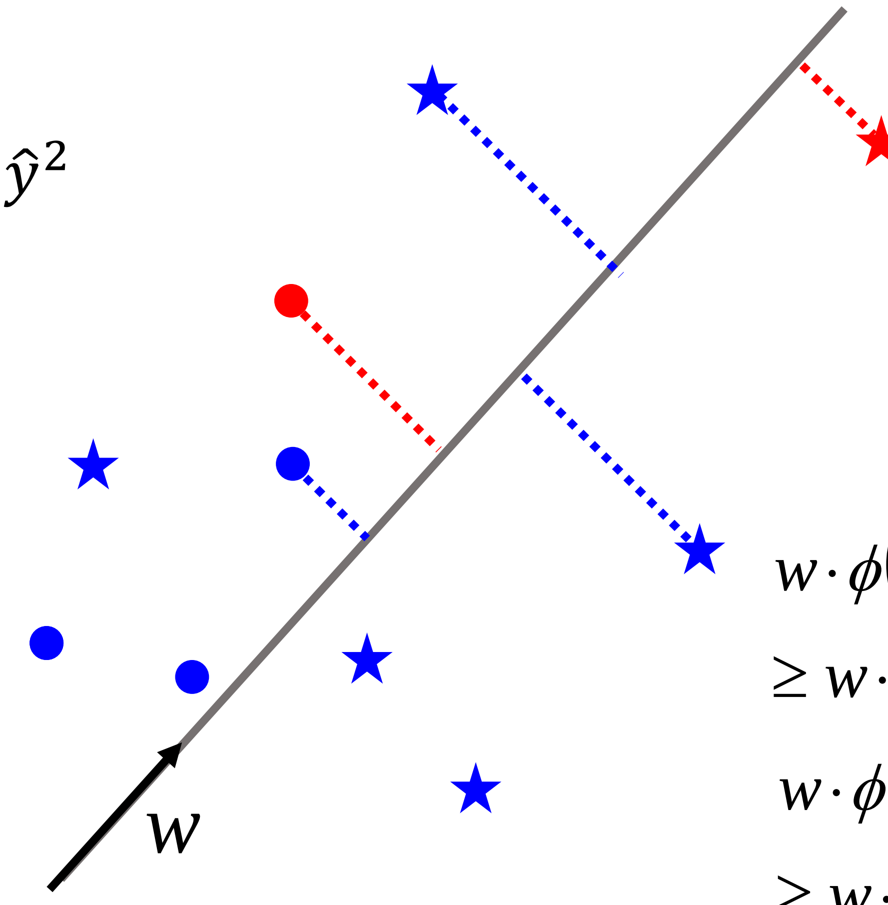
        $$\tilde{y}^r = \arg\max_{y \in Y} w \cdot \phi(x^r, y)$$ (question 2)
      - If $\tilde{y}^r \neq \hat{y}^r$, update w

        $$w \rightarrow w + \phi(x^r, \hat{y}^r) - \phi(x^r, \tilde{y}^r)$$
  - until w is not updated ➡ We are done!

Algorithm - Example

x^1 $\hat{y}^1$ x^2 $\hat{y}^2$

- $\phi(x^1, \hat{y}^1)$ (red circle)
- $\phi(x^1, y)$ (blue circle)
- $\phi(x^2, \hat{y}^2)$ (red star)
- $\phi(x^2, y)$ (blue star)



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

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

Algorithm - Example

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

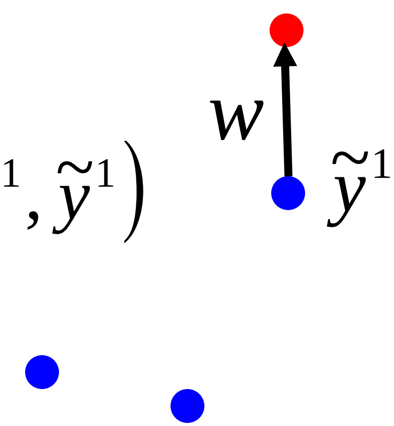
Initialize w = 0

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

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

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

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



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

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

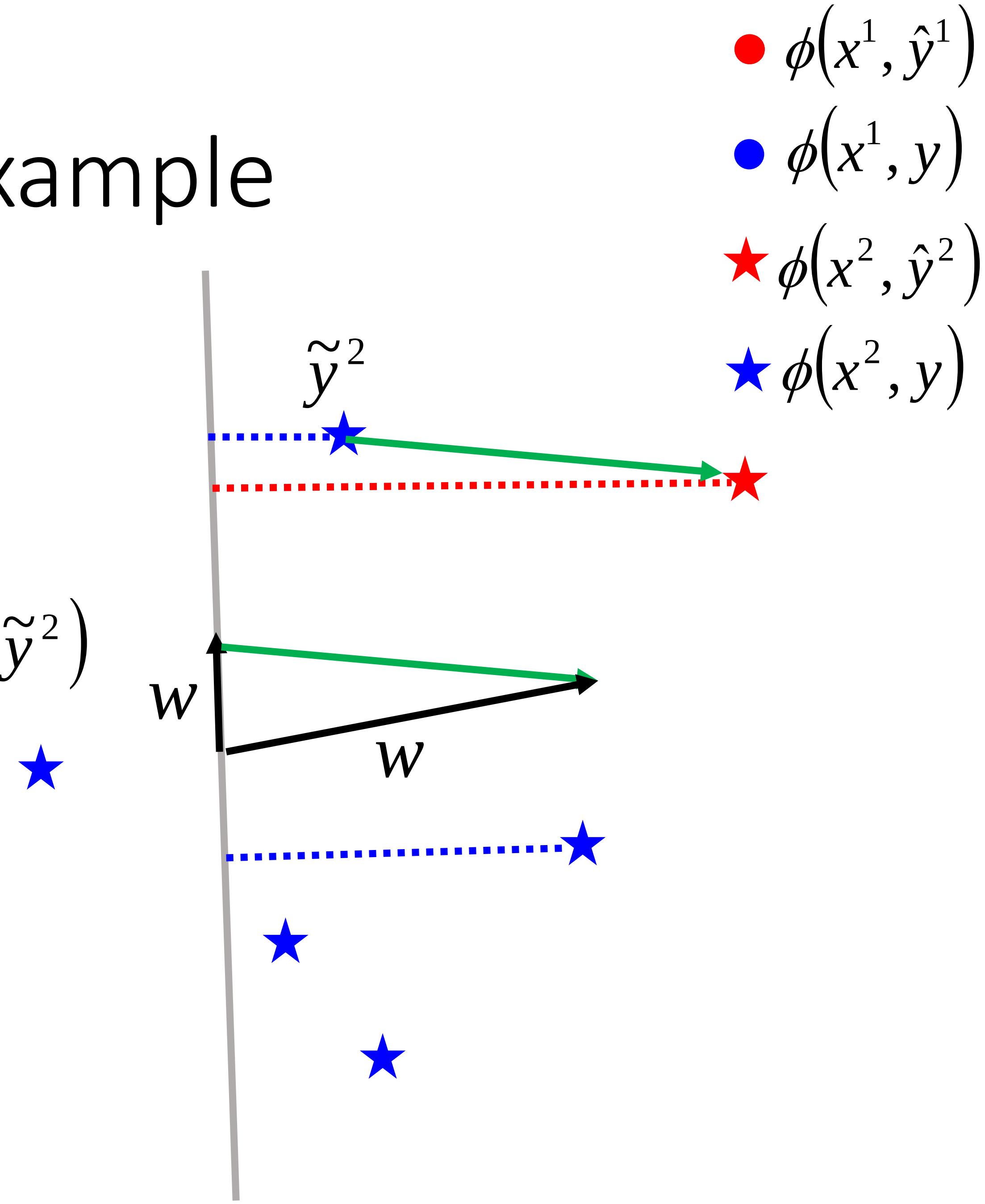
Algorithm - Example

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

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

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

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



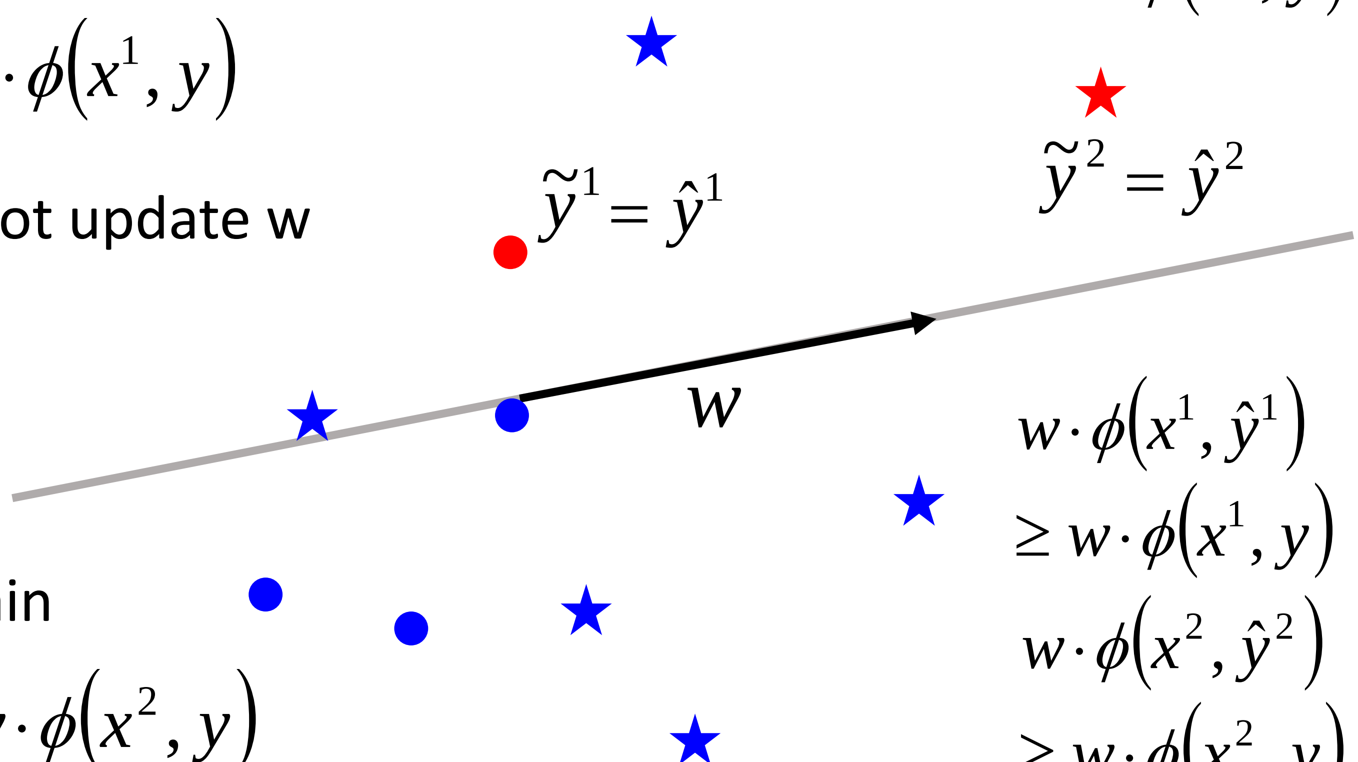
Algorithm - Example

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

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

Assumption: Separable

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

$\forall r$ (All training examples)

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

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

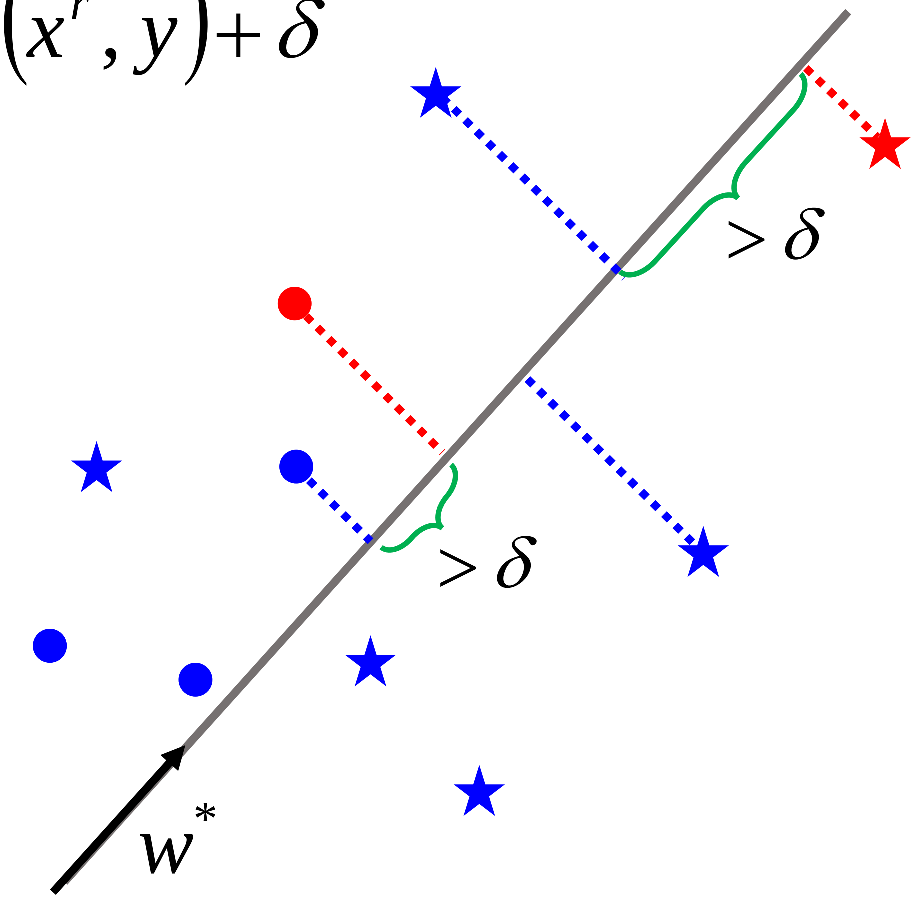
$$\hat{w} \cdot \phi(x^r, \hat{y}^r) \geq \hat{w} \cdot \phi(x^r, y) + \delta$$

Assumption: Separable

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

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

……



Proof of Termination

w is updated once it sees a mistake

$$w^0 = 0 \rightarrow w^1 \rightarrow w^2 \rightarrow \ldots\ldots \rightarrow w^k \rightarrow w^{k+1} \rightarrow \ldots\ldots$$

$$w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)$$ (the relation of w^k and w^{k-1})

Proof that: The angle ρ_k between $\hat{w}$ and w_k is smaller as k increases

Analysis $\cos\rho_k$ (larger and larger?) $\cos\rho_k = \frac{\boxed{\hat{w}}}{\|\hat{w}\|} \cdot \frac{\boxed{w^k}}{\|w^k\|}$

$$\hat{w} \cdot w^k = \hat{w} \cdot \left(w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\right)$$

$$= \hat{w} \cdot w^{k-1} + \underbrace{\hat{w} \cdot \phi(x^n, \hat{y}^n) - \hat{w} \cdot \phi(x^n, \tilde{y}^n)}_{\geq \delta \text{ (Separable)}} \geq \hat{w} \cdot w^{k-1} + \delta$$

Proof of Termination

w is updated once it sees a mistake

$$w^0 = 0 \rightarrow w^1 \rightarrow w^2 \rightarrow \ldots\ldots \rightarrow w^k \rightarrow w^{k+1} \rightarrow \ldots\ldots$$

$$w^k = w^{k-1} + \phi\left(x^n, \hat{y}^n\right) - \phi\left(x^n, \tilde{y}^n\right)$$ (the relation of w^k and w^{k-1})

Proof that: The angle ρ_k between $\hat{w}$ and w_k is smaller as k increases

Analysis $\cos \rho_k$ (larger and larger?) $\cos \rho_k = \dfrac{\boxed{\hat{w}}}{\|\hat{w}\|} \cdot \dfrac{\boxed{w^k}}{\boxed{\|w^k\|}}$

$$\hat{w} \cdot w^k \geq \hat{w} \cdot w^{k-1} + \delta$$

$$\hat{w} \cdot w^1 \geq \underbrace{\hat{w} \cdot w^0}_{=0} + \delta \qquad \hat{w} \cdot w^2 \geq \underbrace{\hat{w} \cdot w^1}_{\geq\delta} + \delta \quad \ldots\ldots$$

$$\hat{w} \cdot w^1 \geq \delta \qquad \hat{w} \cdot w^2 \geq 2\delta \quad \ldots\ldots$$

$$\hat{w} \cdot w^k \geq k\delta$$

(so what)

Proof of Termination

$$\cos\rho_k = \frac{\hat{w}}{\|\hat{w}\|}\cdot\frac{w^k}{\boxed{\|w^k\|}} \qquad w^k = w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)$$

$$\|w^k\|^2 = \|w^{k-1} + \phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\|^2$$

$$= \|w^{k-1}\|^2 + \|\underbrace{\phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)}_{>0}\|^2 + \underbrace{2w^{k-1}\cdot\left(\phi(x^n, \hat{y}^n) - \phi(x^n, \tilde{y}^n)\right)}_{?\ <0\ \text{(mistake)}}$$

Assume the distance between any two feature vector is smaller than R

$$\le \|w^{k-1}\| + R^2$$

$$\|w^1\|^2 \le \|w^0\|^2 + R^2 = R^2$$

$$\|w^2\|^2 \le \|w^1\|^2 + R^2 \le 2R^2$$

$$\cdots$$

$$\|w^k\|^2 \le kR^2$$

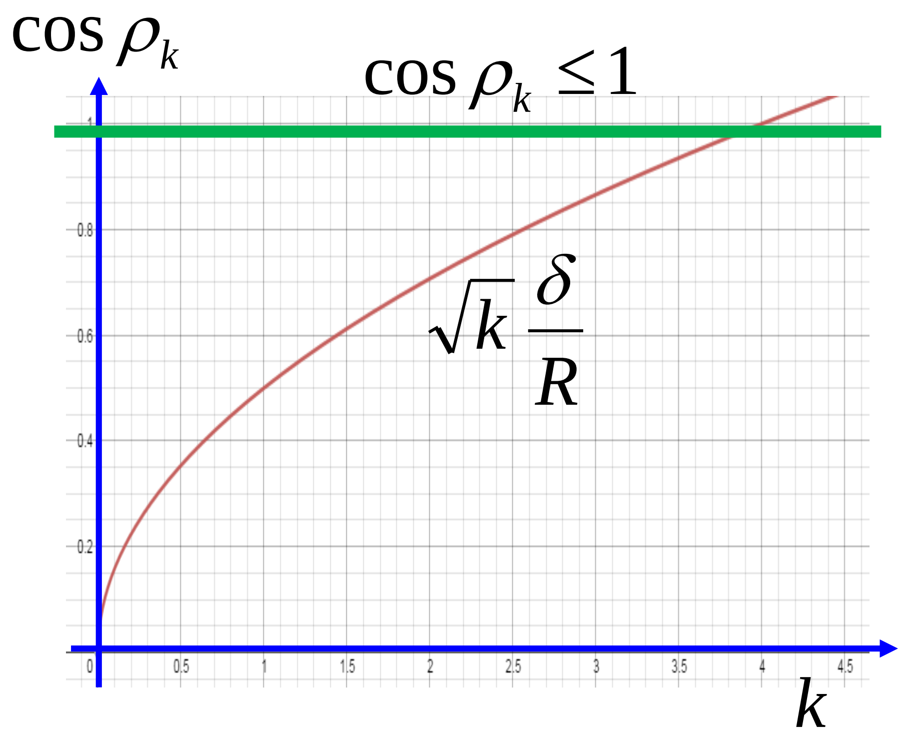
Proof of Termination

$$\cos\rho_k = \frac{\hat{w}}{\|\hat{w}\|}\cdot\frac{w^k}{\|w^k\|} \qquad \hat{w}\cdot w^k \geq k\delta \qquad \|w^k\|^2 \leq k\mathrm{R}^2$$

$$\geq \frac{k\delta}{\sqrt{kR^2}} = \sqrt{k}\,\frac{\delta}{R}$$

$$\sqrt{k}\,\frac{\delta}{R} \leq 1$$

$$k \leq \left(\frac{R}{\delta}\right)^2$$

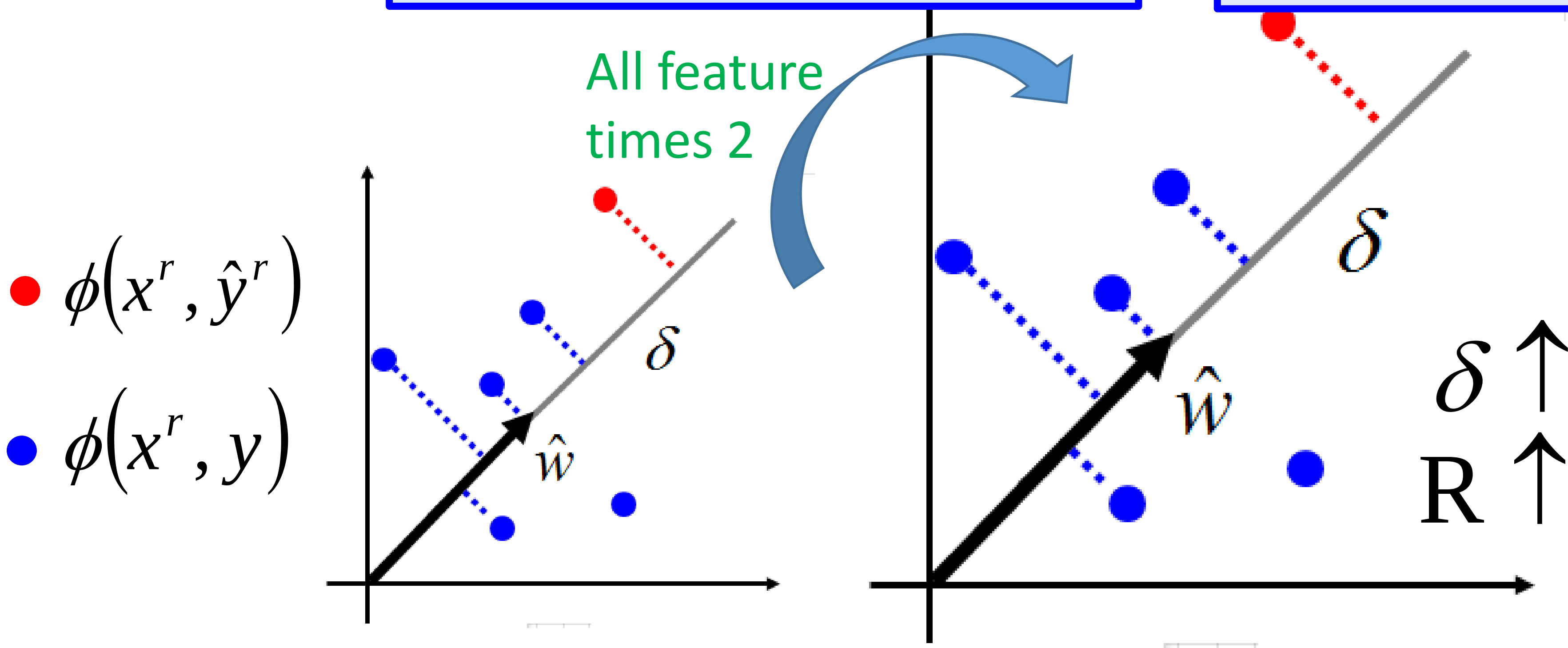


Proof of Termination

$$k \le \left(\frac{R}{\delta}\right)^2$$

The largest distances between features → Normalization

Margin: Is it easy to separable red points from the blue ones → Larger margin, less update



Structured Linear Model: Reduce 3 Problems to 2

Problem 1: Evaluation

- How to define F(x,y)

Problem 2: Inference

- How to find the y with the largest F(x,y)

Problem 3: Training

- How to learn F(x,y)

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

Problem A: Feature

- How to define $\phi(x,y)$

Problem B: Inference

- How to find the y with the largest $w \cdot \phi(x,y)$