Structured Support Vector Machine

Hung-yi Lee

Structured Learning

- We need a more powerful function *f*
  - Input and output are both objects with structures
  - *Object*: sequence, list, tree, bounding box …

$$f : X \rightarrow Y$$

X is the space of one kind of object

Y is the space of another kind of object

Unified Framework

Step 1: Training

- Find a function F

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

- F(x,y): evaluate how compatible the objects x and y is

Step 2: Inference (Testing)

- Given an object x

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

Three Problems

Problem 1: Evaluation

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

Problem 2: Inference

- How to solve the "arg max" problem

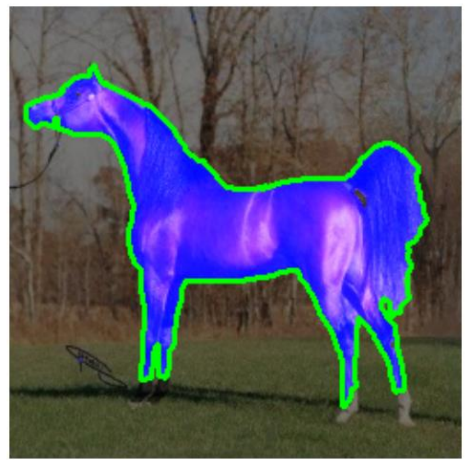
$$y = \arg\max_{y \in Y} F(x, y)$$

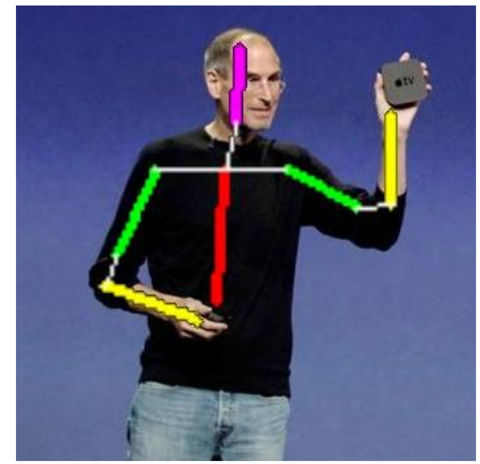
Problem 3: Training

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

Example Task: Object Detection

Example Task





Keep in mind that what you will learn today can be applied to other tasks.

Source of image:
http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.295.6007&rep=rep1&type=pdf
http://www.vision.ee.ethz.ch/~hpedemo/gallery.php

Problem 1: Evaluation

- F(x,y) is linear

$x \rightarrow$  $y \rightarrow$

$F($  $) = w \cdot \phi($  $)$

Open question: What if F(x,y) is not linear?

Problem 2: Inference

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

$w \cdot \phi($  $)$=1.1 …… $w \cdot \phi($  $)$=8.2 ……

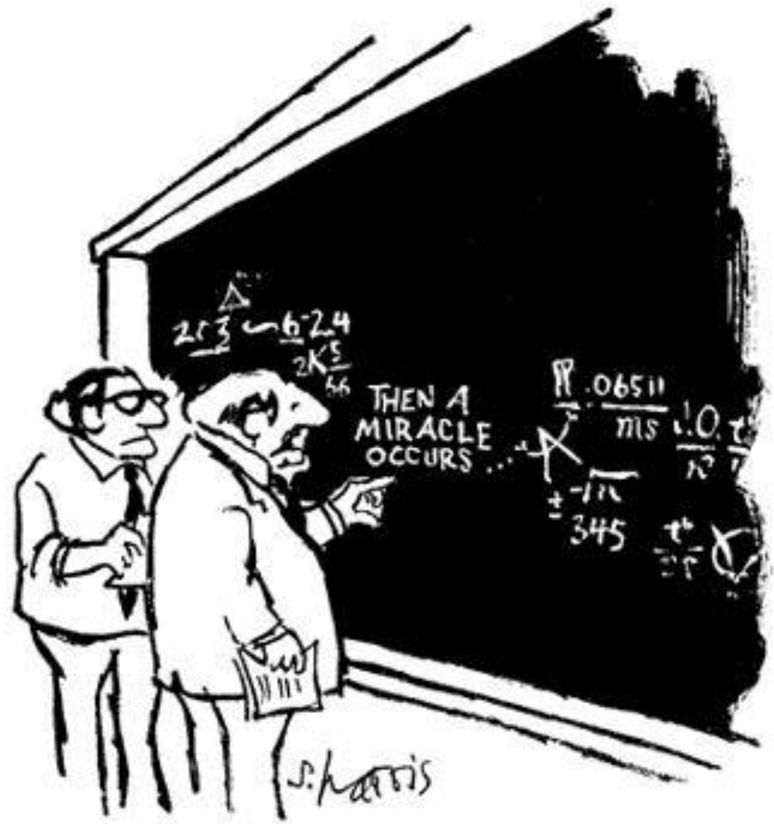
⋮ ⋮

$w \cdot \phi($  $)$=0.3 …… $w \cdot \phi($  **max** $)$=10.1 $\tilde{y}$ ……

⋮ ⋮

$w \cdot \phi($  $)$=-1.5 …… $w \cdot \phi($  $)$ =5.6 ……

Problem 2: Inference



"I think you should be more explicit here in step two."

http://www.condenaststore.com/-sp/I-think-you-should-be-more-explicit-here-in-step-two-Cartoon-Prints_i8562937_.htm

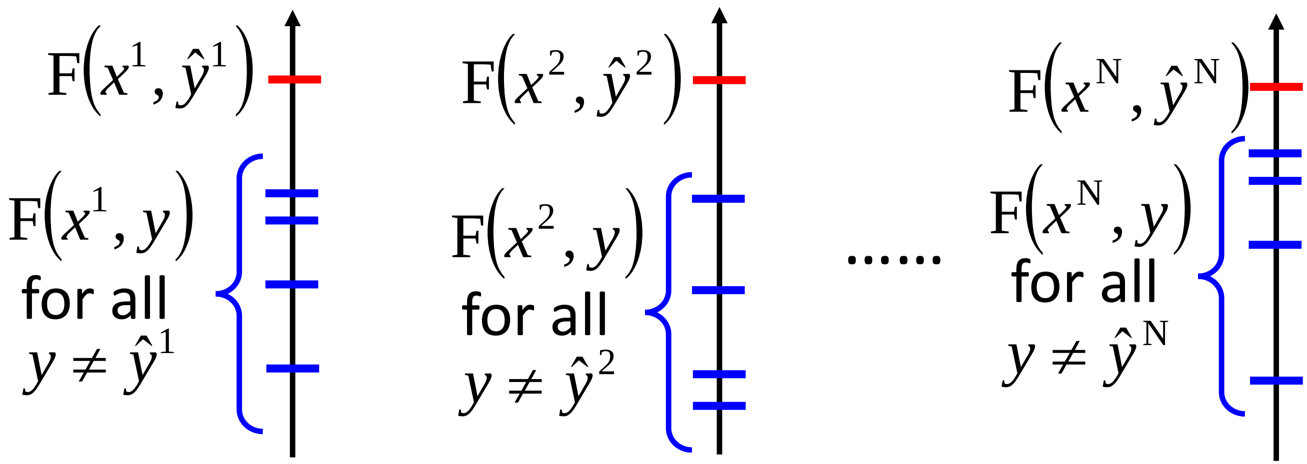
- Object Detection
  - Branch and Bound algorithm
  - Selective Search
- Sequence Labeling
  - Viterbi Algorithm
- The algorithms can depend on $\phi(x, y)$
- Genetic Algorithm
- Open question:
  - What happens if the inference is non exact?

Problem 3: Training

Principle

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

We should find F(x,y) such that ……



Let's ignore problems 1 and 2 and only focus on problem 3 today.

Outline

Separable case

Non-separable case

Considering Errors

Regularization

Structured SVM

Cutting Plane Algorithm for Structured SVM

Multi-class and binary SVM

Beyond Structured SVM (open question)

Outline

Separable case

Non-separable case

Considering Errors

Regularization

Structured SVM

Cutting Plane Algorithm for Structured SVM

Multi-class and binary SVM

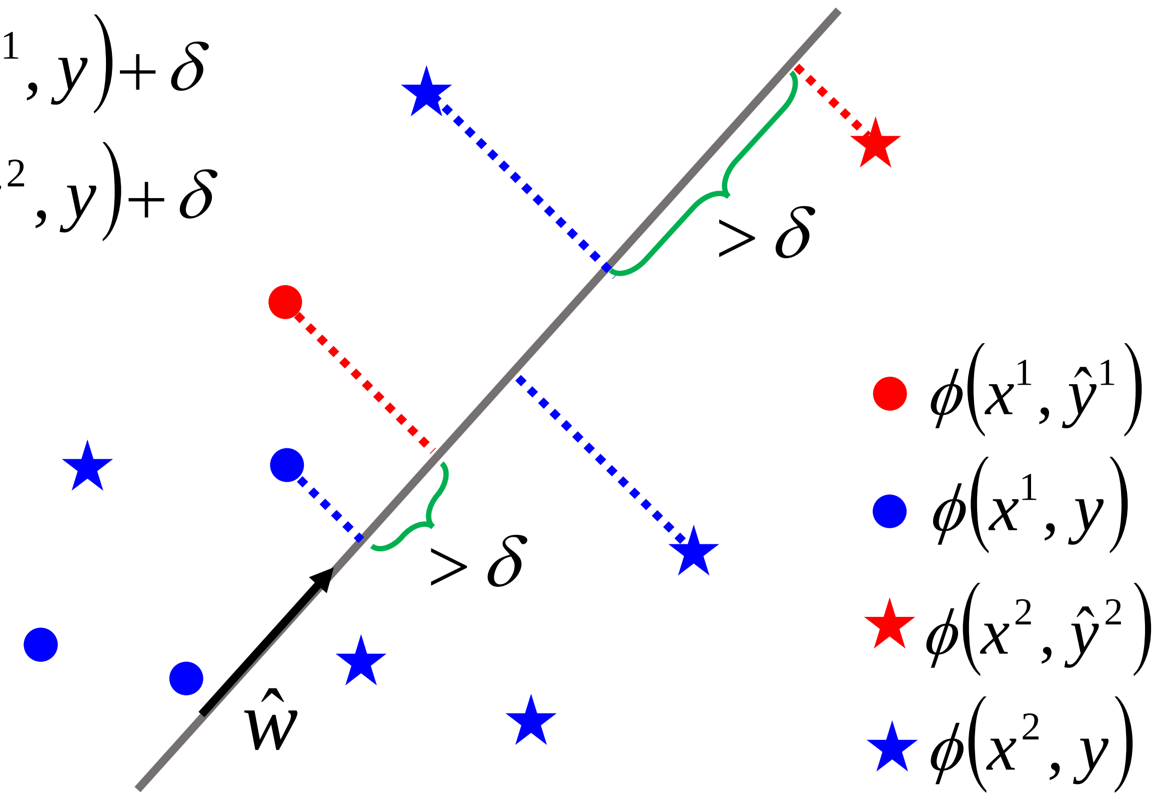
Beyond Structured SVM (open question)

Assumption: Separable

- There exists a weight vector $\hat{w}$

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

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



Structured Perceptron

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

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

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

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

  - until w is not updated $\Rightarrow$ We are done!

Warning of Math

In separable case, to obtain a $\hat{w}$, you only have to update at most $(R/\delta)^2$ times

δ: margin

R: the largest distance between $\phi(x, y)$ and $\phi(x, y')$

Not related to the space of y!

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})

Remind: we are considering the separable case

Assume there exists a weight vector $\widehat{w}$ such that

$\forall n$ (All training examples)

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

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

Assume $\|\widehat{w}\| = 1$ without loss of generality

Proof of Termination

w is updated once it sees a mistake

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

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

$$\cos\rho_k = \boxed{\frac{\hat{w}}{\|\hat{w}\|} \cdot \frac{w^k}{\|w^k\|}}$$

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

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

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?)

$$\cos \rho_k = \frac{\hat{w}}{\|\hat{w}\|} \cdot \frac{w^k}{\boxed{\|w^k\|}}$$

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

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

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

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

(so what)

Proof of Termination

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

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

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

(mistake)

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

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

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

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

$$\cdots$$

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

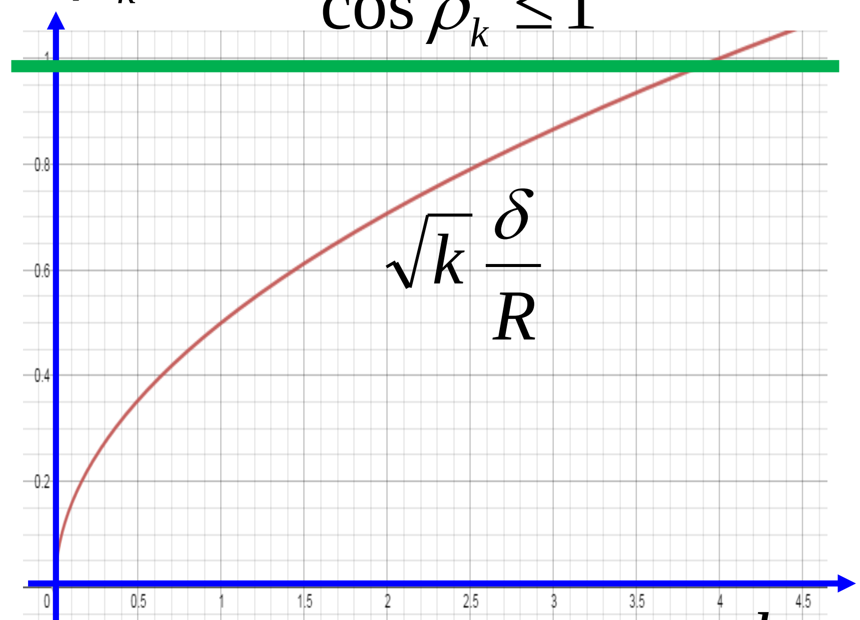
Proof of Termination

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

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

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

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



End of Warning

In separable case, to obtain a $\hat{w}$, you only have to update at most $(R/\delta)^2$ times

δ: margin

R: the largest distance between $\phi(x,y)$ and $\phi(x,y')$

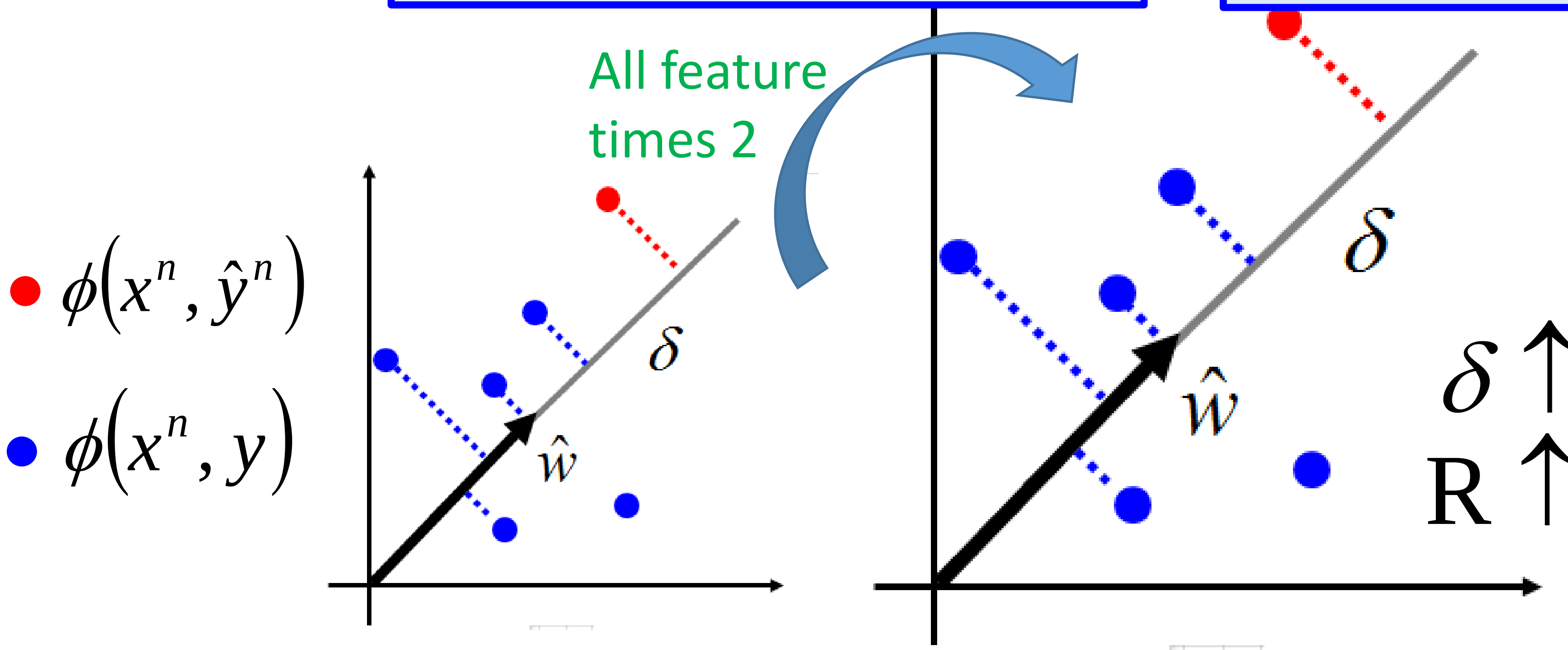
Not related to the space of y!

How to make training fast?

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

The largest distances between features → Normalization

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



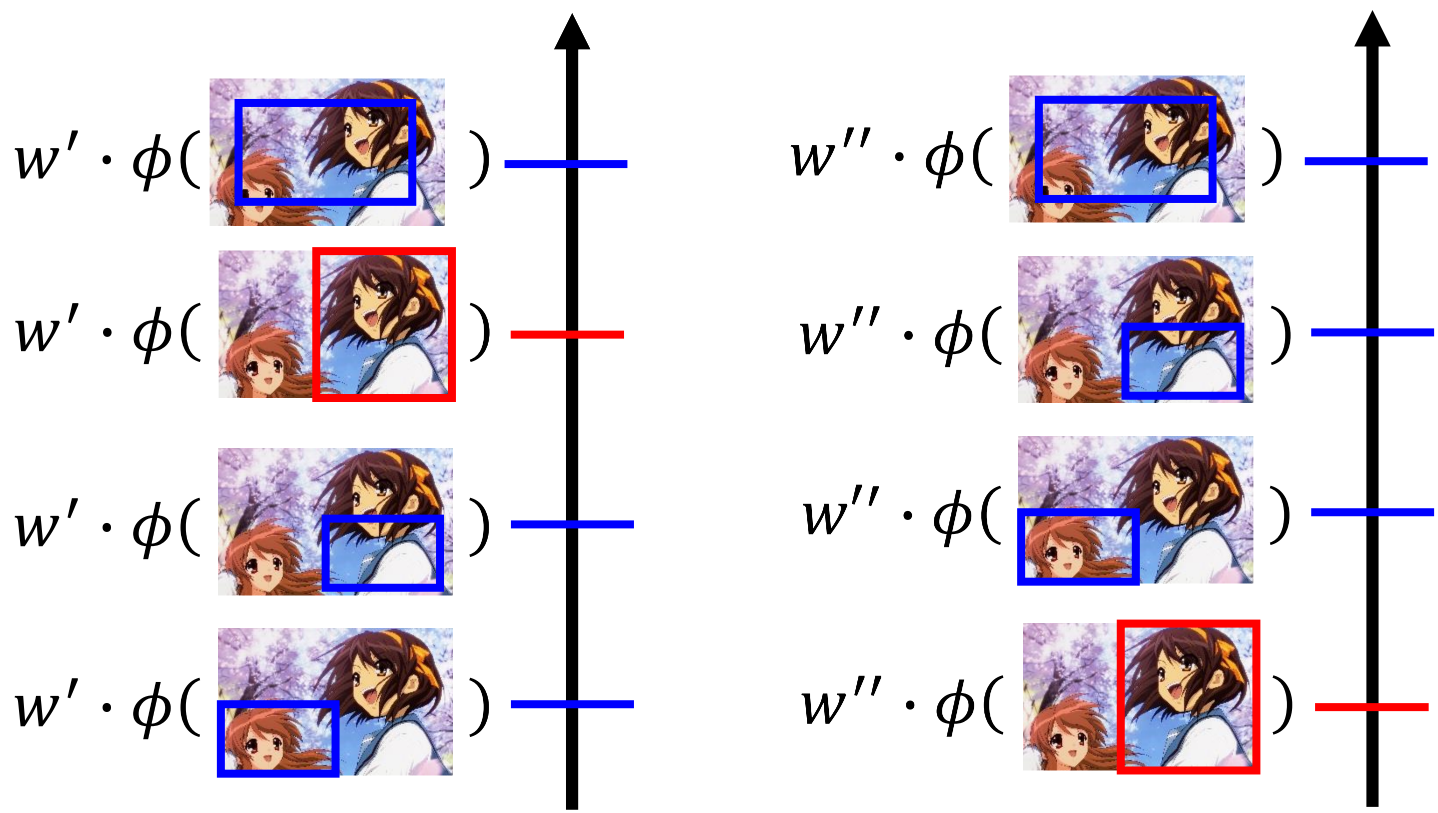
Outline

- Separable case
- Non-separable case
- Considering Errors
- Regularization
- Structured SVM
- Cutting Plane Algorithm for Structured SVM
- Multi-class and binary SVM
- Beyond Structured SVM (open question)

Non-separable Case

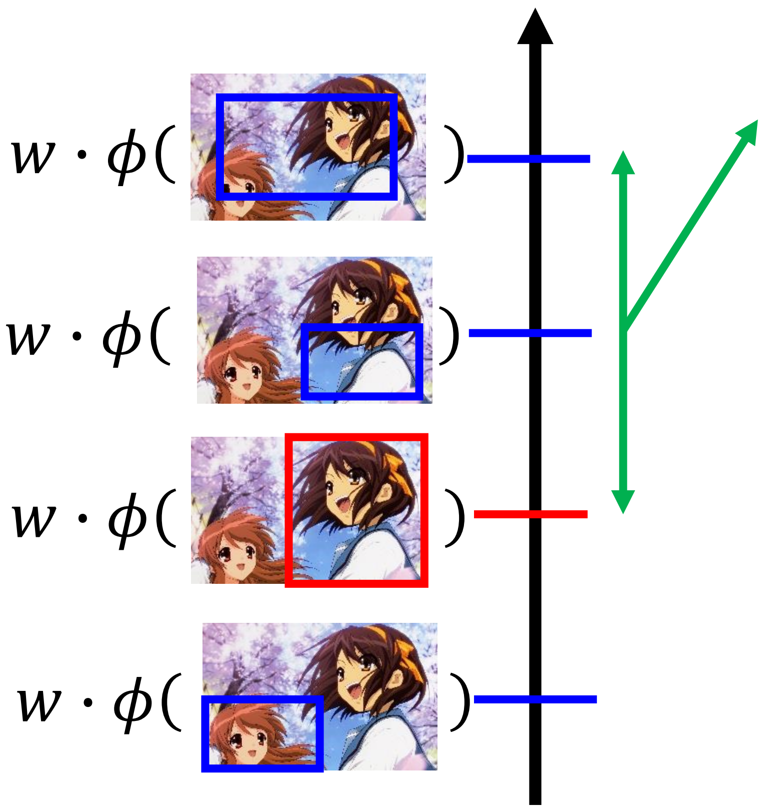
Undoubtedly, w' is better than w''.

- When the data is non-separable, some weights are still better than the others.



Defining Cost Function

- Define a cost C to evaluate how bad a w is, and then pick the w minimizing the cost C



$$C^n = \max_y [w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

$$C = \sum_{n=1}^{N} C^n$$

What is the minimum value?

Other alternatives?

(Stochastic) Gradient Descent

Find w minimizing the cost C

$$C = \sum_{n=1}^{N} C^n$$

$$C^n = \max_y [w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

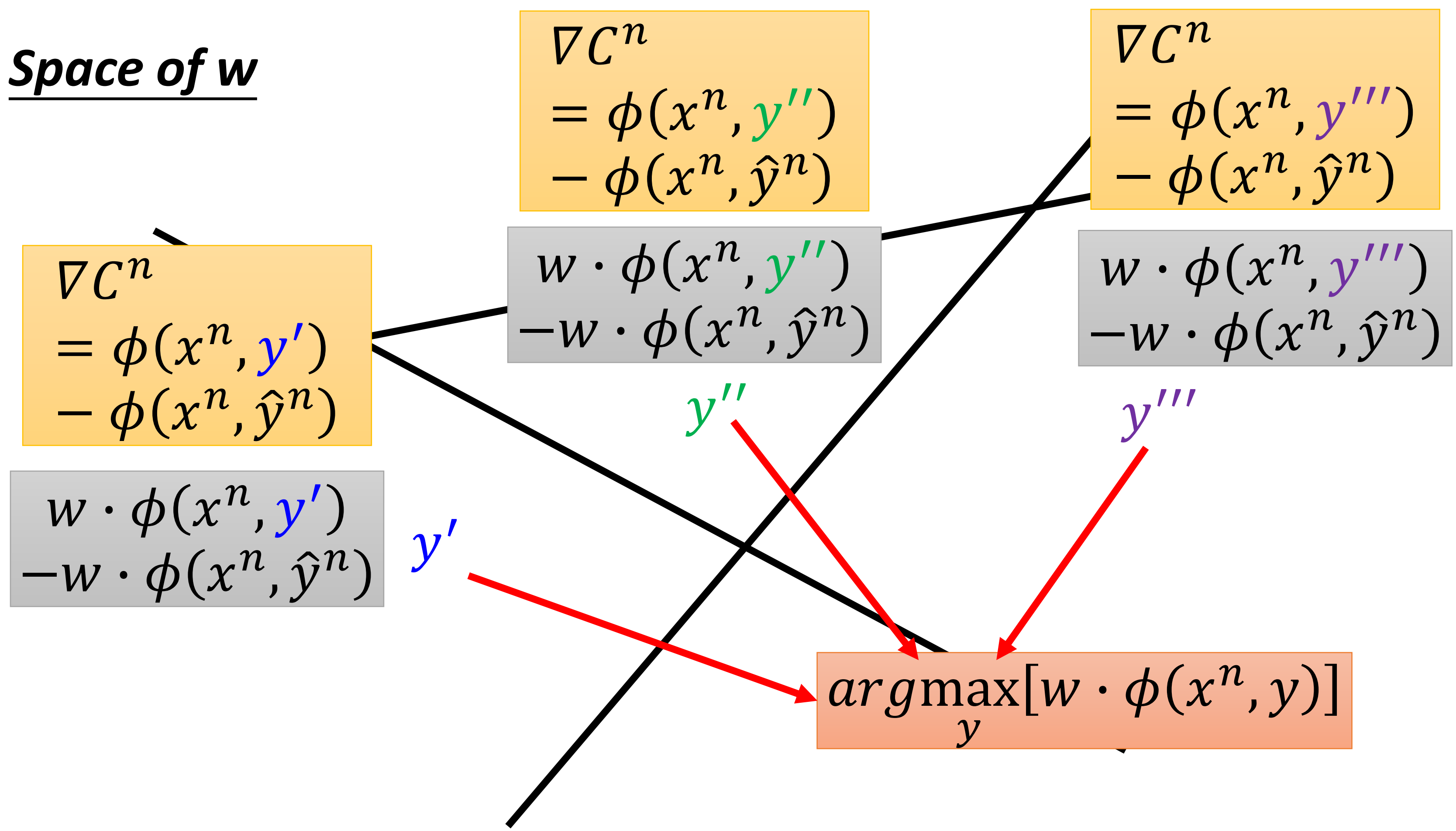
<u>(Stochastic) Gradient descent:</u>

We only have to know how to compute ∇C^n.

However, there is "max" in C^n ....

$$C^n = \underbrace{\max_y [w \cdot \phi(x^n, y)]}_{\text{When w is different, the y can be different.}} - w \cdot \phi(x^n, \hat{y}^n)$$

How to compute ∇C^n?



(Stochastic) Gradient Descent

For t = 1 to T: Update the parameters T times

Randomly pick a training data $\{x^n, \hat{y}^n\}$ stochastic

$$\tilde{y}^n = arg\max_y [w \cdot \phi(x^n, y)]$$ Locate the region

$$\nabla C^n = \phi(x^n, \tilde{y}^n) - \phi(x^n, \hat{y}^n)$$ simple

$$w \rightarrow w - \eta \nabla C^n$$

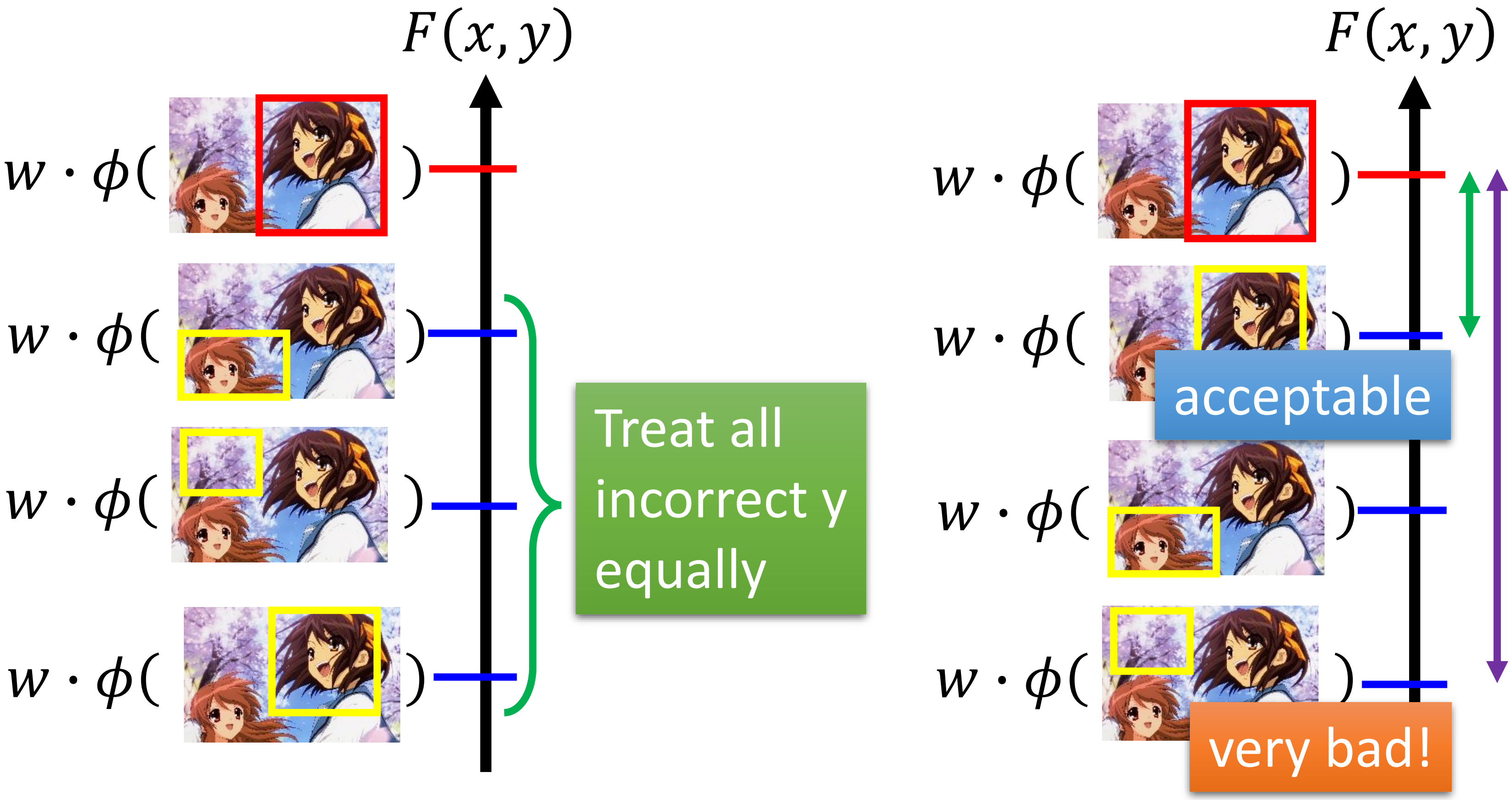
$$= w - \eta[\phi(x^n, \tilde{y}^n) - \phi(x^n, \hat{y}^n)]$$

If we set $\eta = 1$, then we are doing structured perceptron.

Outline

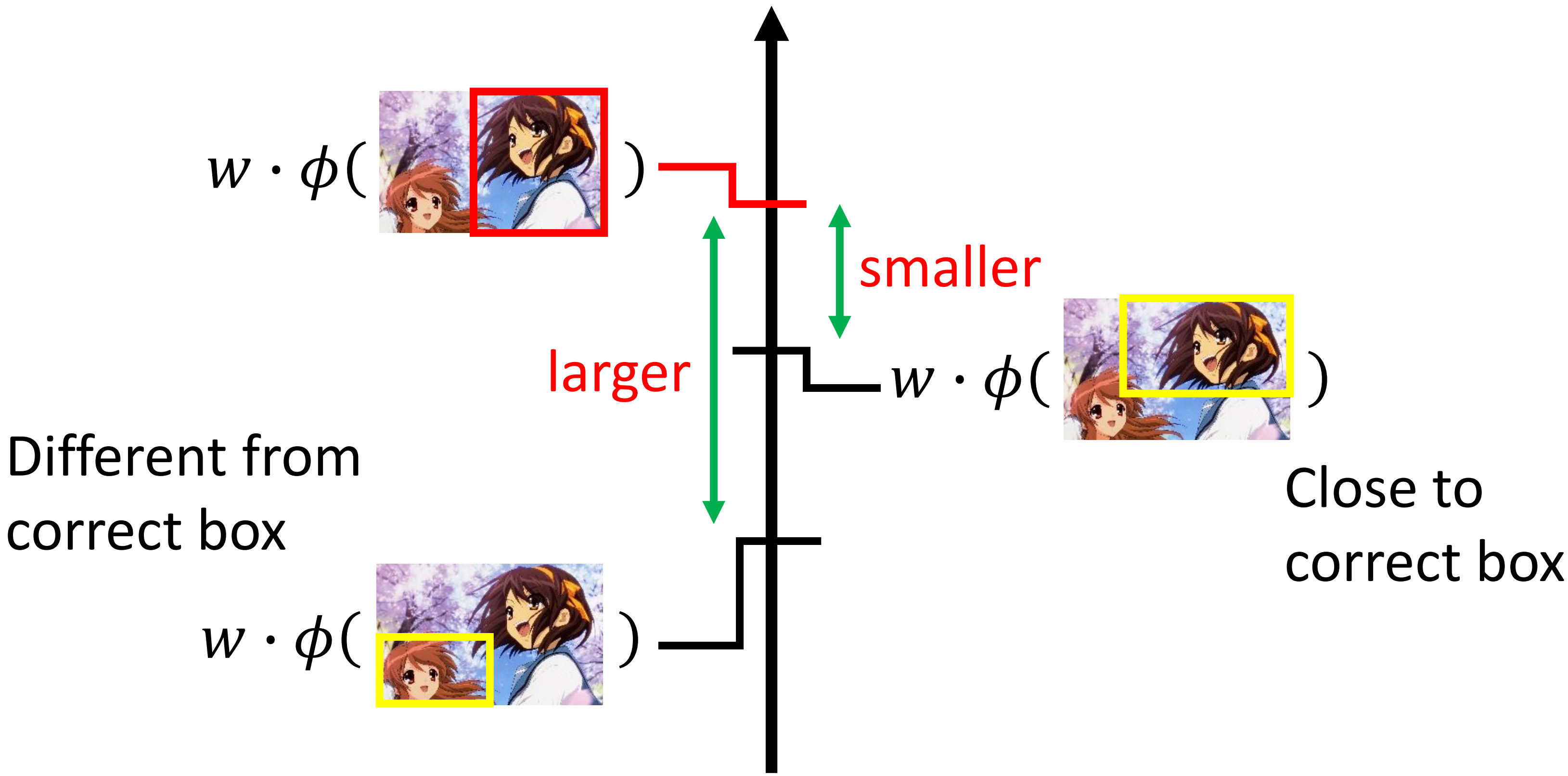
- Separable case
- Non-separable case
- Considering Errors
- Regularization
- Structured SVM
- Cutting Plane Algorithm for Structured SVM
- Multi-class and binary SVM
- Beyond Structured SVM (open question)

Based on what we have considered ……



The right case is better.

Considering the incorrect ones



How to measure the difference

Defining Error Function

- $\Delta(\hat{y}, y)$: difference between $\hat{y}$ and y (> 0)

y

$\hat{y}$

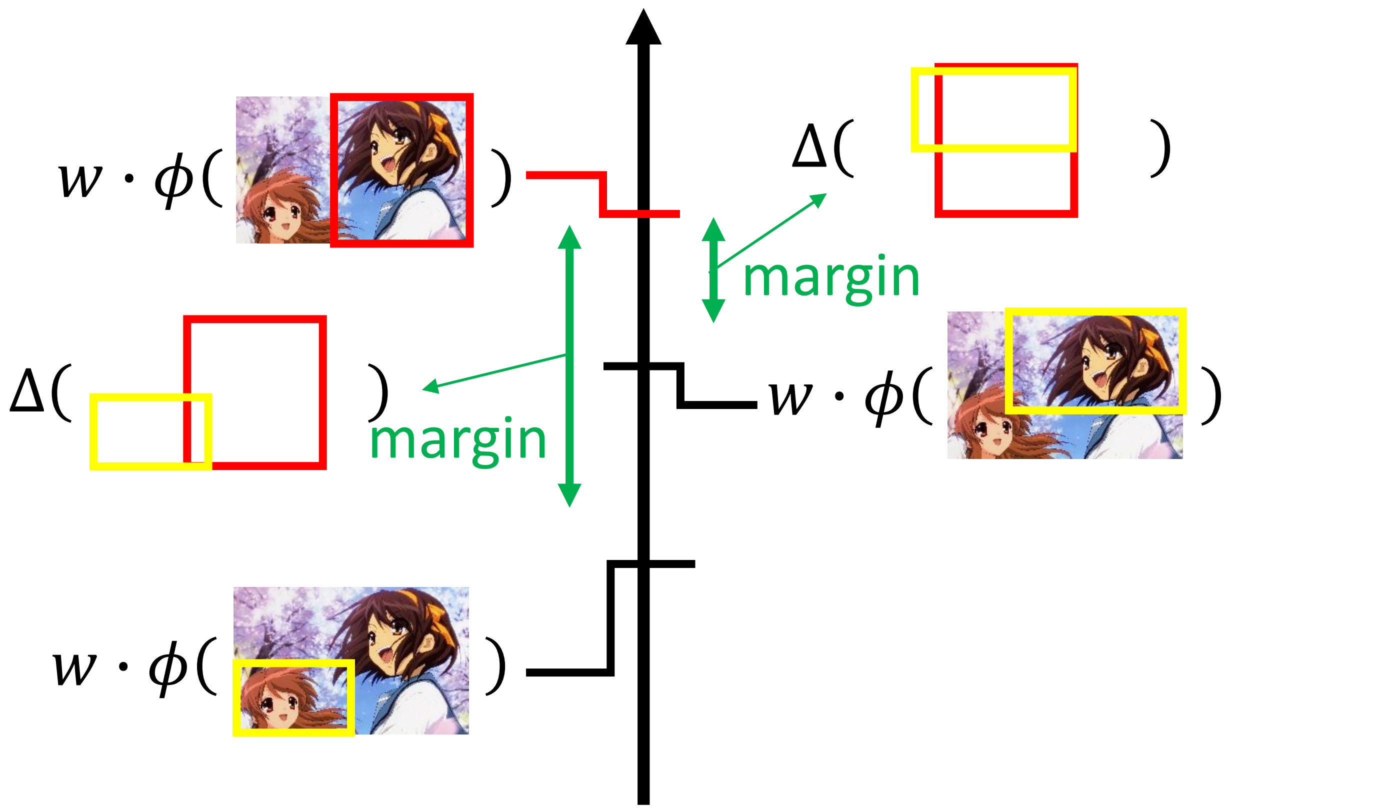
$A(y)$: area of bounding box y

$$\Delta(\hat{y}, y) = 1 - \frac{A(\hat{y}) \cap A(y)}{A(\hat{y}) \cup A(y)}$$

Another Cost Function

$$C^n = \max_y [w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

$$C^n = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$



Gradient Descent

$$C^n = \max_y [w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

$$C^n = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

In each iteration, pick a training data $\{x^n, \hat{y}^n\}$

$$\underset{\bar{y}^n}{\cancel{\tilde{y}^n}} = \cancel{\arg\max_y [w \cdot \phi(x^n, y)]} \arg\max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)]$$

Oh no! Problem 2.1

$$\nabla C^n(w) = \phi(x^n, \underset{\bar{y}^n}{\cancel{\tilde{y}^n}}) - \phi(x^n, \hat{y}^n)$$

$$w \rightarrow w - \eta[\phi(x^n, \underset{\bar{y}^n}{\cancel{\tilde{y}^n}}) - \phi(x^n, \hat{y}^n)]$$

Another Viewpoint

$$\tilde{y}^n = \arg\max_y w \cdot \phi(x^n, y)$$

- Minimizing the new cost function is minimizing the upper bound of the errors on training set

$$C' = \sum_{n=1}^{N} \Delta(\hat{y}^n, \tilde{y}^n) \quad \leq \quad C = \sum_{n=1}^{N} C^n$$

upper bound

We want to find w minimizing C' (errors)

It is hard!

Because y can be any kind of objects, $\Delta(\cdot,\cdot)$ can be any function ……

C serves as the surrogate of C'

Proof that $\Delta(\hat{y}^n, \tilde{y}^n) \leq C^n$

Another Viewpoint

$$C^n = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

Proof that $\Delta(\hat{y}^n, \tilde{y}^n) \leq C^n$

$$\Delta(\hat{y}^n, \tilde{y}^n) \leq \Delta(\hat{y}^n, \tilde{y}^n) + \underbrace{[w \cdot \phi(x^n, \tilde{y}^n) - w \cdot \phi(x^n, \hat{y}^n)]}_{\geq 0}$$

$$\tilde{y}^n = \arg\max_y w \cdot \phi(x^n, y)$$

$$= [\Delta(\hat{y}^n, \tilde{y}^n) + w \cdot \varphi(x^n, \tilde{y}^n)] - w \cdot \varphi(x^n, \hat{y}^n)$$

$$\leq \max_y [\Delta(\hat{y}^n, y) + w \cdot \varphi(x^n, y)] - w \cdot \varphi(x^n, \hat{y}^n)$$

$$= C^n$$

More Cost Functions

$$\Delta(\hat{y}^n, \tilde{y}^n) \le C^n$$

<u>Margin rescaling:</u>

$$C^n = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

<u>Slack variable rescaling:</u>

$$C^n = \max_y \Delta(\hat{y}^n, y)[1 + w \cdot \phi(x^n, y) - w \cdot \phi(x^n, \hat{y}^n)]$$

Outline

- Separable case
- Non-separable case
- Considering Errors
- Regularization
- Structured SVM
- Cutting Plane Algorithm for Structured SVM
- Multi-class and binary SVM
- Beyond Structured SVM (open question)

Regularization

Training data and testing data can have different distribution.

w close to zero can minimize the influence of mismatch.

$$C = \sum_{n=1}^{N} C^n$$

$$C^n = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

$$C = \underbrace{\frac{1}{2}\|w\|^2}_{\text{Regularization: Find the w close to zero}} + \lambda \underbrace{\sum_{n=1}^{N} C^n}_{\text{Keep the incorrect answer from a margin depending on errors}}$$

Keep the incorrect answer from a margin depending on errors

Regularization:
Find the w close to zero

Regularization

$$C = \sum_{n=1}^{N} C^n \Rightarrow C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} C^n$$

In each iteration, pick a training data $\{x^n, \hat{y}^n\}$

$$\bar{y}^n = \arg\max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)]$$

$$\nabla C^n = \phi(x^n, \bar{y}^n) - \phi(x^n, \hat{y}^n) + w$$

$$w \rightarrow w - \eta[\phi(x^n, \bar{y}^n) - \phi(x^n, \hat{y}^n)] - \eta w$$

$$= (1 - \eta)w - \eta[\phi(x^n, \bar{y}^n) - \phi(x^n, \hat{y}^n)]$$

Weight decay as in DNN

Outline

- Separable case
- Non-separable case
- Considering Errors
- Regularization
- Structured SVM
- Cutting Plane Algorithm for Structured SVM
- Multi-class and binary SVM
- Beyond Structured SVM (open question)

Structured SVM

Find w minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} C^n$$

$$C^n = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

$$C^n + w \cdot \phi(x^n, \hat{y}^n) = \max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)]$$

Are they equivalent?

We want to minimize C

For $\forall y$:

$$C^n + w \cdot \phi(x^n, \hat{y}^n) \geq \Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)$$

$$w \cdot \phi(x^n, \hat{y}^n) - w \cdot \phi(x^n, y) \geq \Delta(\hat{y}^n, y) - C^n$$

Structured SVM

Find w minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} C^n$$

$$C^n = \max_y[\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)] - w \cdot \phi(x^n, \hat{y}^n)$$

$$\equiv$$

Find w,$\varepsilon^1, \cdots, \varepsilon^N$ minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \boxed{\varepsilon^n}$$

For $\forall n$:

For $\forall y$:

$$w \cdot \phi(x^n, \hat{y}^n) - w \cdot \phi(x^n, y) \geq \Delta(\hat{y}^n, y) - \boxed{\varepsilon^n}$$

Slack variable

Structured SVM

Find w,$\varepsilon^1, \cdots, \varepsilon^N$ minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \boxed{\varepsilon^n}$$

For $\forall n$:

For $\forall y$:

$$w \cdot \phi(x^n, \hat{y}^n) - w \cdot \phi(x^n, y) \geq \Delta(\hat{y}^n, y) - \boxed{\varepsilon^n}$$

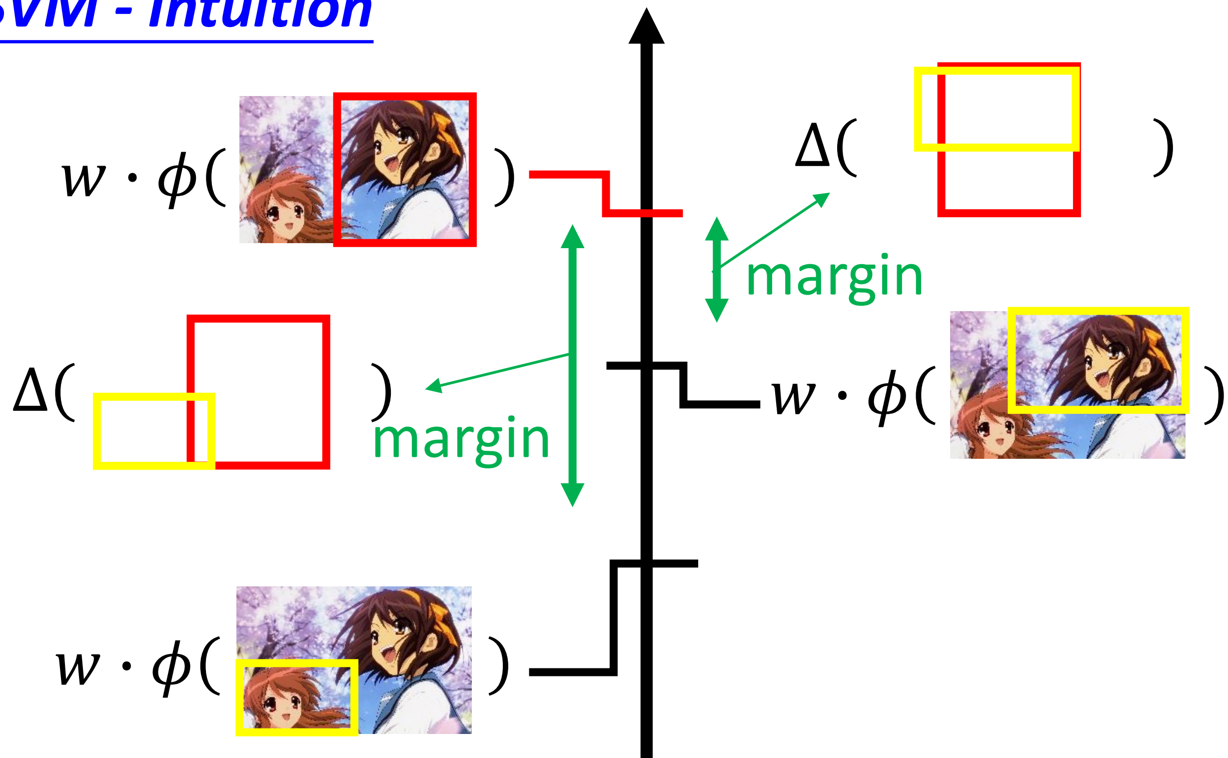
For $\forall y \neq \hat{y}^n$:

$$w \cdot \left(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\right) \geq \Delta(\hat{y}^n, y) - \varepsilon^n, \quad \varepsilon^n \geq 0$$

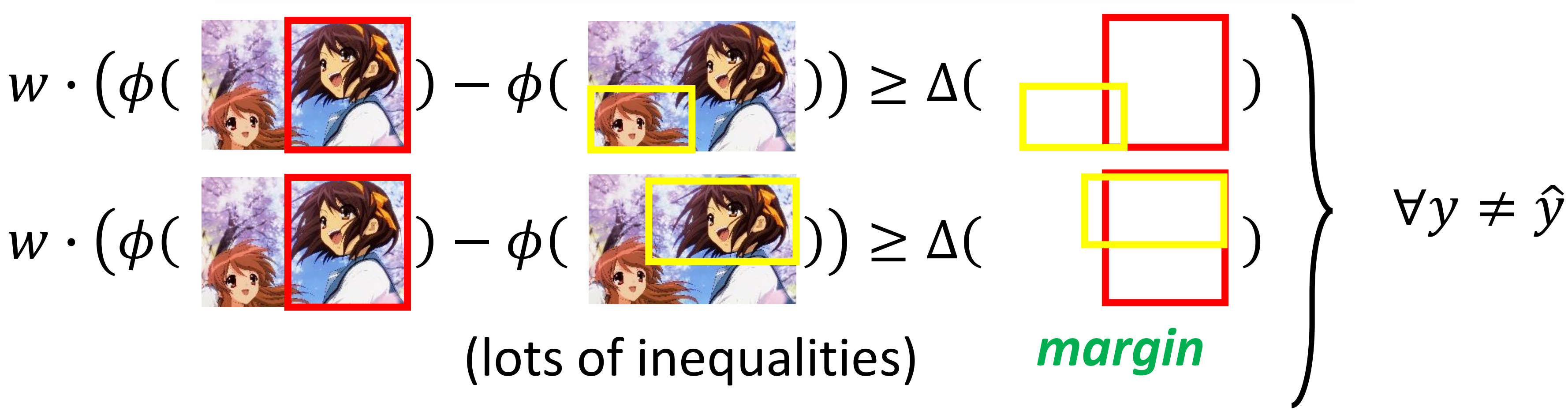
If y $= \hat{y}^n$: $\underbrace{w \cdot \phi(x^n, \hat{y}^n) - w \cdot \phi(x^n, \hat{y}^n)}_{=0} \geq \underbrace{\Delta(\hat{y}^n, \hat{y}^n)}_{=0} - \varepsilon^n$

$\Rightarrow \varepsilon^n \geq 0$

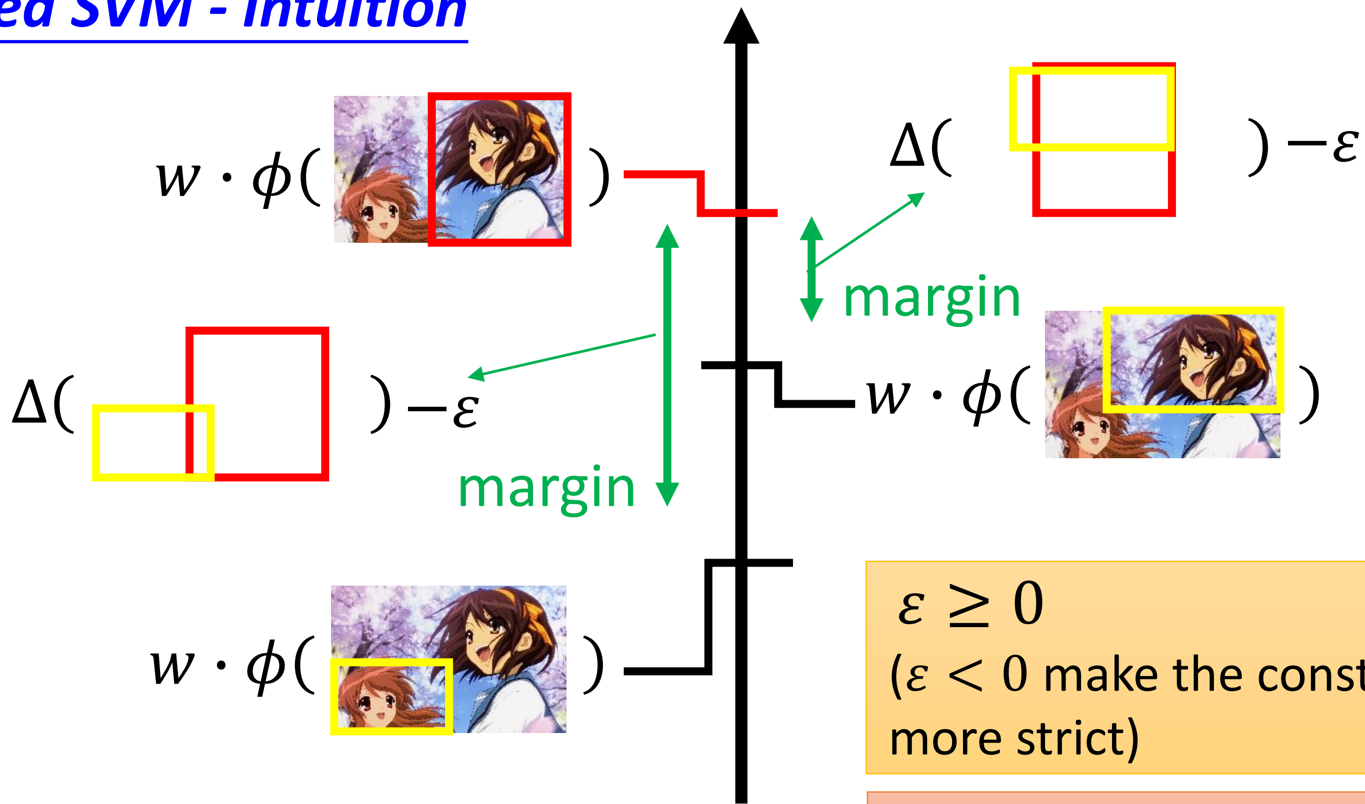
Structured SVM - Intuition



It is possible that no w can achieve this.



Structured SVM - Intuition



$\varepsilon \geq 0$
($\varepsilon < 0$ make the constraints more strict)

ε should be minimized

(lots of inequalities)

slack variable

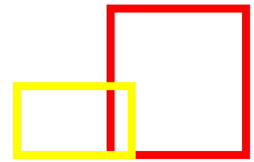
Structured SVM - Intuition

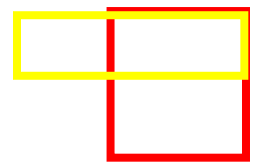
Training data: $\hat{y}^1$ $\hat{y}^2$

x^1  x^2 

Minimize $\frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{2} \varepsilon^n$

For x^1

$w \cdot ($ $\phi($  $) - \phi($  $)) \geq \Delta($  $) - \varepsilon^1$

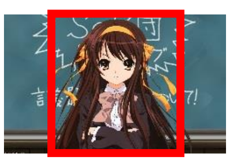
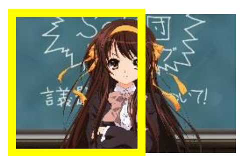
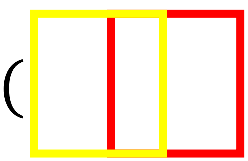
$w \cdot ($ $\phi($  $) - \phi($  $)) \geq \Delta($  $) - \varepsilon^1$

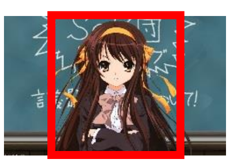
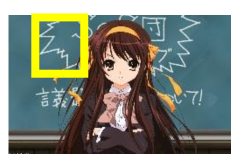
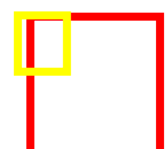
$\forall y \neq \hat{y}^1$

(lots of inequalities)

$\varepsilon^1 \geq 0$

For x^2

$w \cdot ($ $\phi($  $) - \phi($  $)) \geq \Delta($  $) - \varepsilon^2$

$w \cdot ($ $\phi($  $) - \phi($  $)) \geq \Delta($  $) - \varepsilon^2$

$\forall y \neq \hat{y}^2$

(lots of inequalities)

$\varepsilon^2 \geq 0$

Structured SVM

Find w,$\varepsilon^1, \cdots, \varepsilon^N$ minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \boxed{\varepsilon^n}$$

For $\forall n$:

For $\forall y \neq \hat{y}^n$:

$$w \cdot \left(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\right) \geq \Delta(\hat{y}^n, y) - \varepsilon^n, \quad \varepsilon^n \geq 0$$

Solve it by the solver in SVM package

Quadratic Programming (QP) Problem

Too many constraints ……

Outline

- Separable case
- Non-separable case
- Considering Errors
- Regularization
- Structured SVM
- Cutting Plane Algorithm for Structured SVM
- Multi-class and binary SVM
- Beyond Structured SVM (open question)

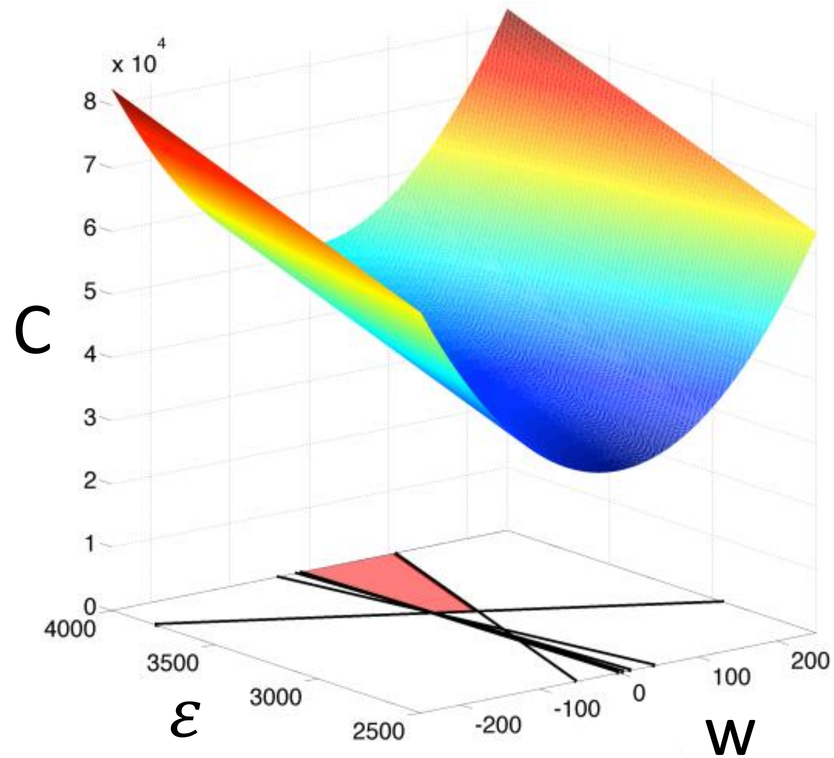
Find w,$\varepsilon^1, \cdots, \varepsilon^N$ minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \boxed{\varepsilon^n}$$

For $\forall n$:

For $\forall y \neq \hat{y}^n$:

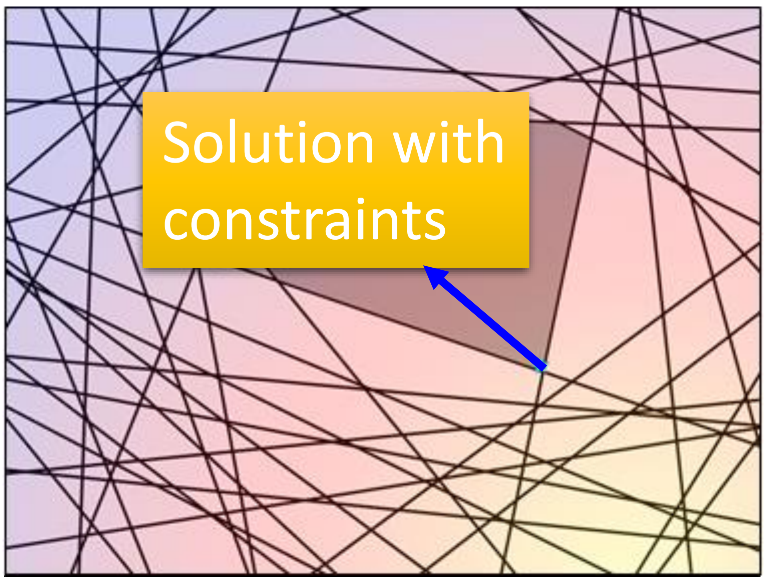
$$w \cdot \left(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\right) \geq \Delta(\hat{y}^n, y) - \varepsilon^n, \ \varepsilon^n \geq 0$$



Source of image: http://abnerguzman.com/publications/gkb_aistats13.pdf

Image credit: Yisong Yue

Cutting Plane Algorithm



Parameter space

$(w, \varepsilon^1, \ldots \varepsilon^N)$

Color is the value of C which is going to be minimized:

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \varepsilon^n$$

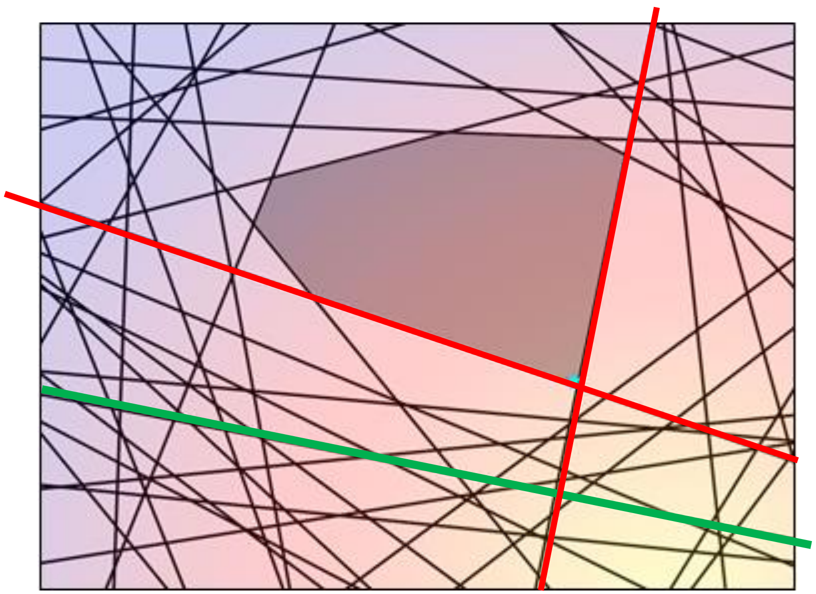
For $\forall r, \forall y, y \neq \hat{y}^n$:

- $w \cdot \left(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\right) \geq \Delta(\hat{y}^n, y) - \varepsilon^n$
- $\varepsilon^n \geq 0$

Image credit: Yisong Yue

Cutting Plane Algorithm

Although there are lots of constraints, most of them do not influence the solution.



Parameter space

$(w, \varepsilon^1, \dots, \varepsilon^N)$

Red lines: determine the solution

Green line: Remove this constraint will not influence the solution

$y \in \mathbb{A}^n$

For $\forall r, \cancel{\forall y}, y \neq \hat{y}^n$:

- $w \cdot \left(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\right) \geq \Delta(\hat{y}^n, y) - \varepsilon^n$
- $\varepsilon^n \geq 0$

$\mathbb{A}^n$: a very small set of y →***working set***

Cutting Plane Algorithm

- Elements in ***working set*** $\mathbb{A}^n$ is selected iteratively

Initialize $\mathbb{A}^1 \ldots \mathbb{A}^N$

Find $w, \varepsilon^1 \ldots \varepsilon^N$ minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \varepsilon^n$$

Solve a QP problem

For $\forall r$:

For $\forall y \in \mathbb{A}^n, y \neq \hat{y}^n$:

$$w \cdot \left(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\right) \geq \Delta(\hat{y}^n, y) - \varepsilon^n$$

$$\varepsilon^n \geq 0$$

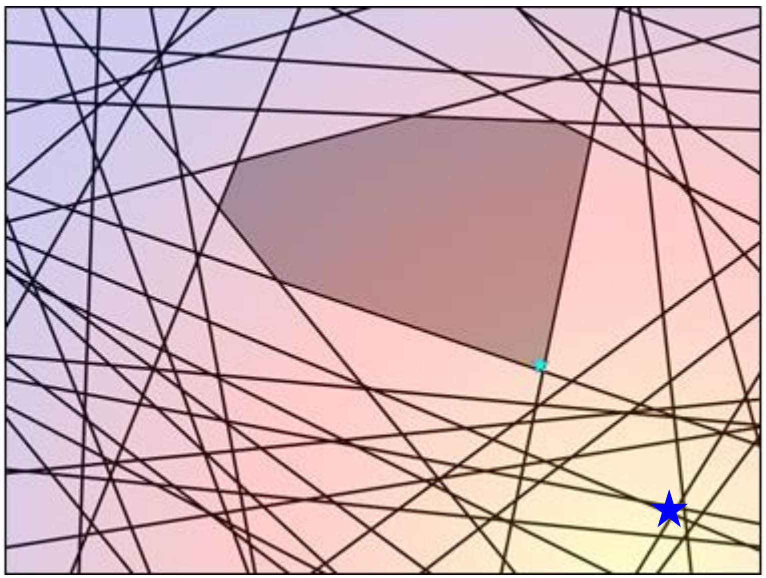
obtain solution w

Repeatedly

Add elements into $\mathbb{A}^1 \ldots \mathbb{A}^N$

Cutting Plane Algorithm

- Strategies of adding elements into ***working set*** $\mathbb{A}^n$



Initialize $\mathbb{A}^n = null$

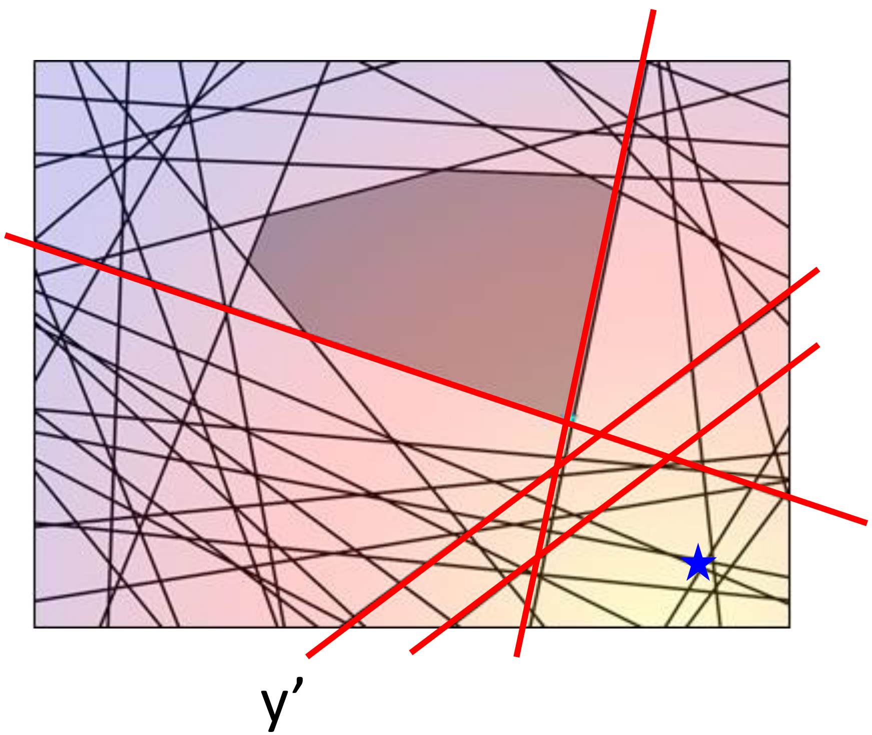
No constraint at all

Solving QP

The solution w is the blue point.

Cutting Plane Algorithm

- Strategies of adding elements into ***working set*** $\mathbb{A}^n$



There are lots of constraints is violated

Find ***<u>the most violated one</u>***

Suppose it is the constraint from y’

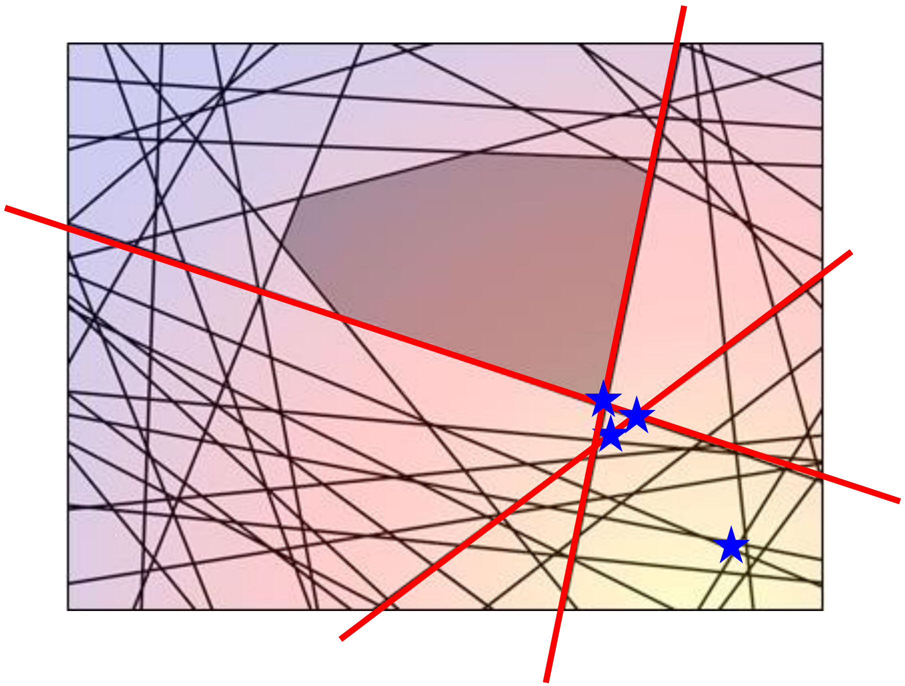
Extent the working set

$$\mathbb{A}^n = \mathbb{A}^n \cup \{y'\}$$

Image credit: Yisong Yue

Cutting Plane Algorithm

- Strategies of adding elements into ***working set*** $\mathbb{A}^n$



Find the most violated one

- Given w' and ε' from working sets at hand, which constraint is the most violated one?

<u>Constraint:</u> $w \cdot \left(\phi(x, \hat{y}) - \phi(x, y)\right) \geq \Delta(\hat{y}, y) - \varepsilon$

<u>Violate a Constraint:</u>

$$w' \cdot \left(\phi(x, \hat{y}) - \phi(x, y)\right) < \Delta(\hat{y}, y) - \varepsilon'$$

<u>Degree of Violation</u>

$$\Delta(\hat{y}, y) - \varepsilon' - w' \cdot \left(\phi(x, \hat{y}) - \phi(x, y)\right)$$

$$\Rightarrow \Delta(\hat{y}, y) + w' \cdot \phi(x, y)$$

<u>The most violated one:</u>

$$\arg\max_{y} [\Delta(\hat{y}, y) + w \cdot \phi(x, y)]$$

Cutting Plane Algorithm

Given training data: $\{(x^1, \hat{y}^1), (x^2, \hat{y}^2), \cdots, (x^N, \hat{y}^N)\}$

Working Set $\mathbb{A}^1 \leftarrow null, \mathbb{A}^2 \leftarrow null, \cdots, \mathbb{A}^N \leftarrow null$

Repeat

w ←Solve a **QP** with Working Set $\mathbb{A}^1, \mathbb{A}^2, \cdots, \mathbb{A}^N$

QP: Find w, $\varepsilon^1 \dots \varepsilon^N$ minimizing $\frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \varepsilon^n$

For $\forall n$:

For $\forall y \in \mathbb{A}^n$:

$$w \cdot \big(\phi(x^n, \hat{y}^n) - \phi(x^n, y)\big) \geq \Delta(\hat{y}^n, y) - \varepsilon^n, \varepsilon^n \geq 0$$

Cutting Plane Algorithm

Given training data: $\{(x^1,\hat{y}^1),(x^2,\hat{y}^2),\cdots,(x^N,\hat{y}^N)\}$

Working Set $\mathbb{A}^1 \leftarrow null, \mathbb{A}^2 \leftarrow null, \cdots, \mathbb{A}^N \leftarrow null$

Repeat

w ←Solve a **QP** with Working Set $\mathbb{A}^1, \mathbb{A}^2, \cdots, \mathbb{A}^N$

For each training data $(x^n,\hat{y}^n)$:

$$\bar{y}^n = \arg\max_y [\Delta(\hat{y}^n, y) + w \cdot \phi(x^n, y)]$$

find the most violated constraints

Update working set $\mathbb{A}^n \leftarrow \mathbb{A}^n \cup \{\bar{y}^n\}$

Until $\mathbb{A}^1, \mathbb{A}^2, \cdots, \mathbb{A}^N$ doesn't change any more

Return w

Training data:

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

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

$\mathbb{A}^1=\{\}$

$\mathbb{A}^2=\{\}$

$w = 0$

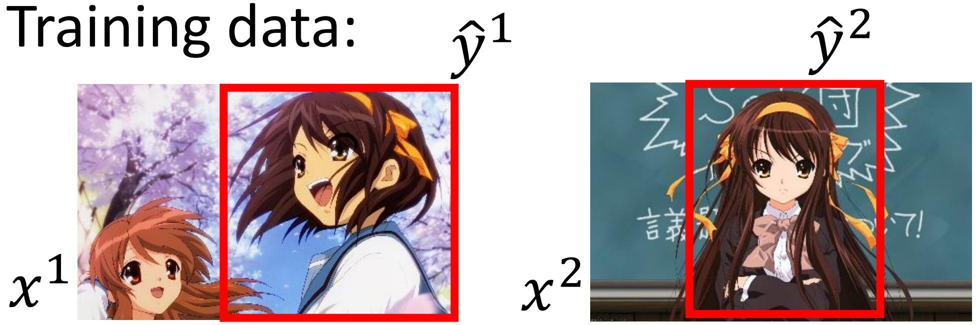
QP: Find w, ε^1,ε^2 minimizing

$$\frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{2} \varepsilon^n$$

There is no constraint

Solution: $w = 0$

Training data:

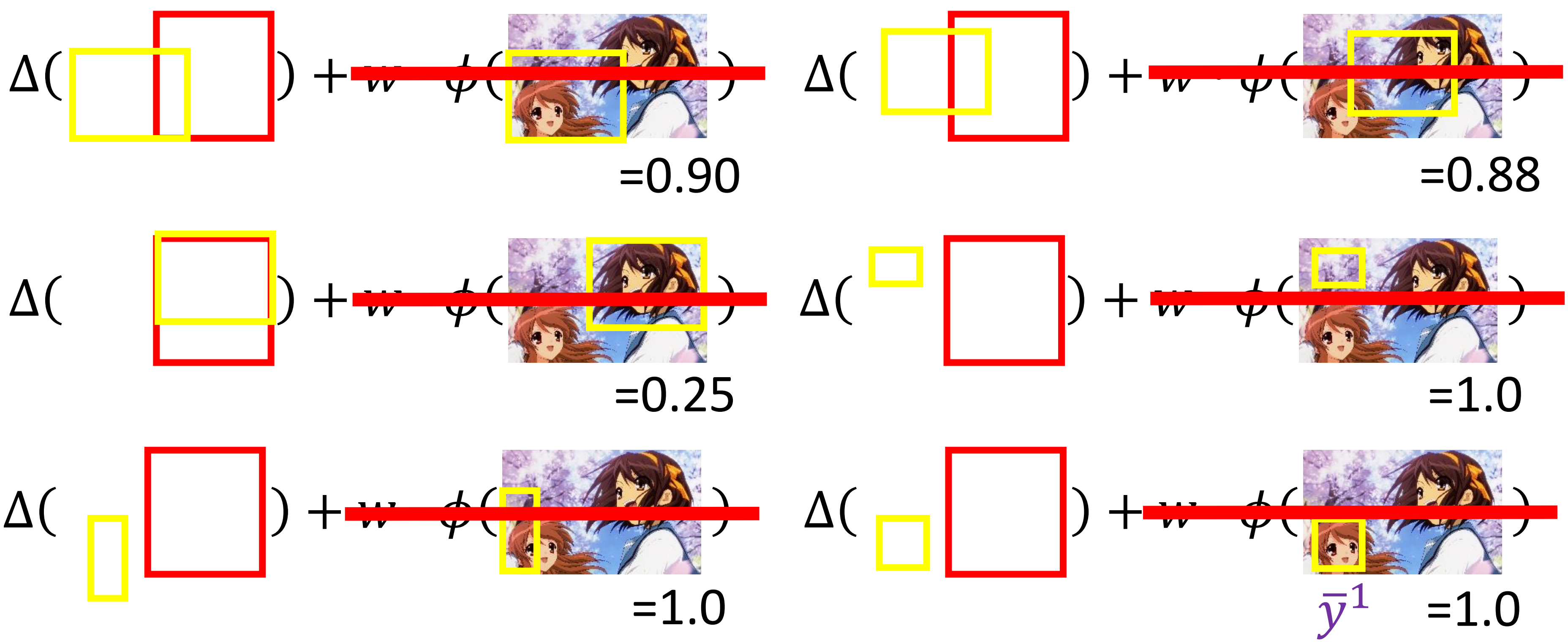


$\mathbb{A}^1=\{\} \longrightarrow \mathbb{A}^1=\{\quad\}$

$\mathbb{A}^2=\{\} \longrightarrow \mathbb{A}^2=\{\quad\}$

$w=0$

$$\bar{y}^1 = \arg\max_y [\Delta(\hat{y}^1, y) + 0 \cdot \phi(x^1, y)]$$



$$\bar{y}^2 = \arg\max_y [\Delta(\hat{y}^2, y) + 0 \cdot \phi(x^2, y)]$$

Training data:

$\hat{y}^1$

x^1 

$\hat{y}^2$

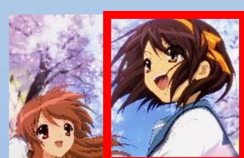
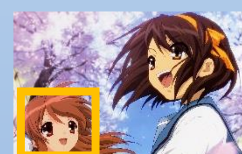
x^2 

$\mathbb{A}^1=\{ \quad \}$

$\mathbb{A}^2=\{ \quad \}$

$w = w^1$

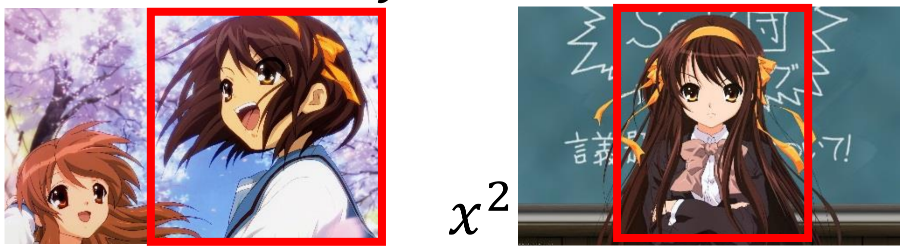
QP: Find w, ε^1,ε^2 minimizing $\frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{2} \varepsilon^n$

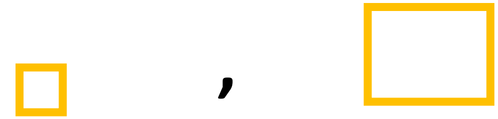
$w \cdot \big(\phi($  $) - \phi($  $)\big) \geq \Delta(\quad) - \varepsilon^1$

$w \cdot \big(\phi($  $) - \phi($  $)\big) \geq \Delta(\quad) - \varepsilon^2$

Solution: $w = w^1$

Training data:

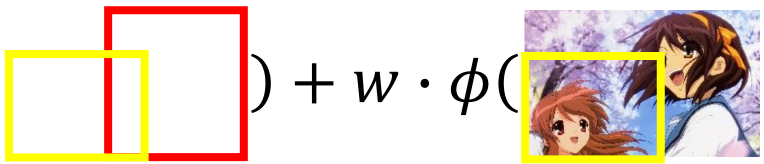


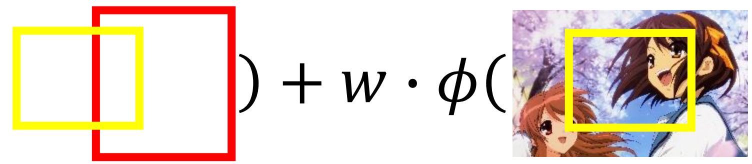
$\mathbb{A}^1$={  }

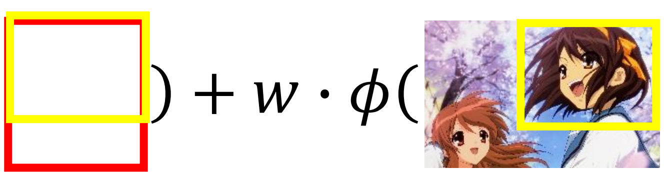
$\mathbb{A}^2$={  }

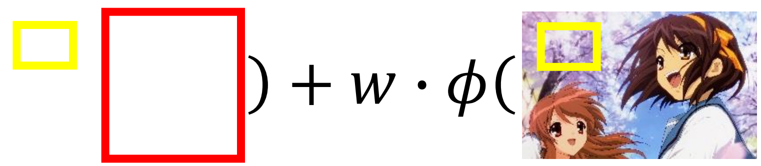
$w = w^1$

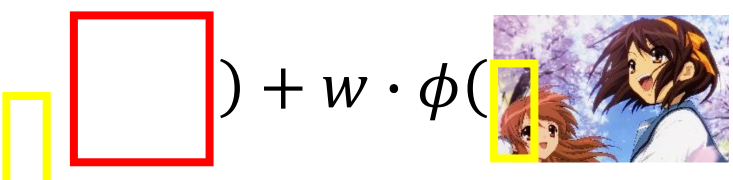
$$\bar{y}^1 = \arg\max_y [\Delta(\hat{y}^1, y) + w^1 \cdot \phi(x^1, y)]$$

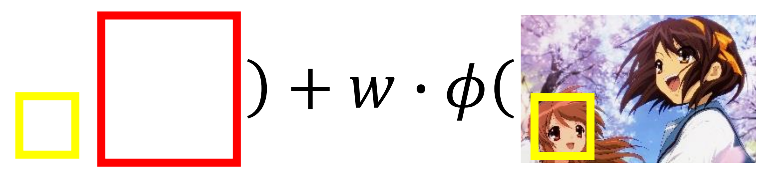
$\Delta($  $)$ =0.97

$\Delta($  $)$ $\bar{y}^1$ =1.55

$\Delta($  $)$ =1.25

$\Delta($  $)$ =1.01

$\Delta($  $)$ =-0.99

$\Delta($  $)$ =-1.10

$$\bar{y}^2 = \arg\max_y [\Delta(\hat{y}^2, y) + w^1 \cdot \phi(x^2, y)]$$

Training data: $\hat{y}^1$ $\hat{y}^2$

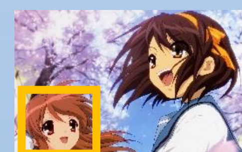
x^1  x^2 

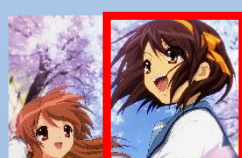
$\mathbb{A}^1 = \{\ ,\ \}$

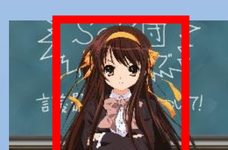
$\mathbb{A}^2 = \{\ ,\ \}$

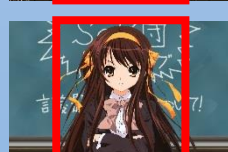
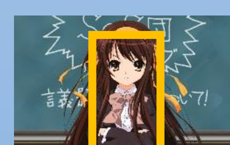
The process repeats iteratively

QP: Find w, $\varepsilon^1, \varepsilon^2$ minimizing $\frac{1}{2}\|w\|^2 + \lambda \sum_{r=1}^{2} \varepsilon^n$

$w \cdot \big(\phi($  $) - \phi($  $)\big) \geq \Delta(\ \ ) - \varepsilon^1$

$w \cdot \big(\phi($  $) - \phi($  $)\big) \geq \Delta(\ \ ) - \varepsilon^1$

$w \cdot \big(\phi($  $) - \phi($  $)\big) \geq \Delta(\ \ ) - \varepsilon^2$

$w \cdot \big(\phi($  $) - \phi($  $)\big) \geq \Delta(\ \ ) - \varepsilon^2$

Concluding Remarks

Separable case

↓

Non-separable case

↓

Considering Errors

↓

Regularization

↓

Structured SVM

↓

Cutting Plane Algorithm for Structured SVM

↓

Multi-class and binary SVM

↓

Beyond Structured SVM (open question)

Multi-class SVM

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

- Problem 1: Evaluation
  - If there are K classes, then we have K weight vectors $\{w^1, w^2, \cdots, w^K\}$

$$y \in \{1,2, \cdots, k, \cdots, K\}$$

$$F(x, y) = w^y \cdot \vec{x}$$

$\vec{x}$: vector representation of x

$$w = \begin{bmatrix} w^1 \\ w^2 \\ \vdots \\ w^k \\ \vdots \\ w^K \end{bmatrix} \quad \phi(x, y) = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \vec{x} \\ \vdots \\ 0 \end{bmatrix}$$

Multi-class SVM

- Problem 2: Inference

$$F(x, y) = w^y \cdot \vec{x}$$

$$\hat{y} = arg \max_{y \epsilon \{1,2,\cdots,k,\cdots,K\}} F(x, y)$$

$$= arg \max_{y \epsilon \{1,2,\cdots,k,\cdots,K\}} w^y \cdot \vec{x}$$

The number of classes are usually small, so we can just enumerate them.

Multi-class SVM

$y \in \{dog, cat, bus, car\}$

$\Delta(\hat{y}^n = dog, y = cat) = 1$

$\Delta(\hat{y}^n = dog, y = bus) = 100$

(defined as your wish)

- Problem 3: Training

Find w,$\varepsilon^1, \cdots, \varepsilon^N$ minimizing C

$$C = \frac{1}{2}\|w\|^2 + \lambda \sum_{n=1}^{N} \varepsilon^n$$

For $\forall n$:

For $\forall y \neq \hat{y}^n$:

There are only N(K-1) constraints.

$$\left(w^{\hat{y}^n} - w^y\right) \cdot \vec{x} \geq \underline{\Delta(\hat{y}^n, y)} - \varepsilon^n, \ \varepsilon^n \geq 0$$

$$w \cdot \phi(x^n, \hat{y}^n) = w^{\hat{y}^n} \cdot \vec{x}$$

$$w \cdot \phi(x^n, y) = w^y \cdot \vec{x}$$

Some types of misclassifications may be worse than others.

Binary SVM

- Set K = 2 $\quad y \in \{1,2\}$

For $\forall y \neq \hat{y}^n$:

$$\left(w^{\hat{y}^n} - w^{y}\right) \cdot \vec{x} \geq \underbrace{\Delta(\hat{y}^n, y)}_{=1} - \varepsilon^n, \ \varepsilon^n \geq 0$$

If y=1: $(w^1 - w^2) \cdot \vec{x} \geq 1 - \varepsilon^n \Rightarrow w \cdot \vec{x} \geq 1 - \varepsilon^n$

w

If y=2: $(w^2 - w^1) \cdot \vec{x} \geq 1 - \varepsilon^n \Rightarrow -w \cdot \vec{x} \geq 1 - \varepsilon^n$

$-w$

Concluding Remarks

Separable case

Non-separable case

Considering Errors

Regularization

Structured SVM

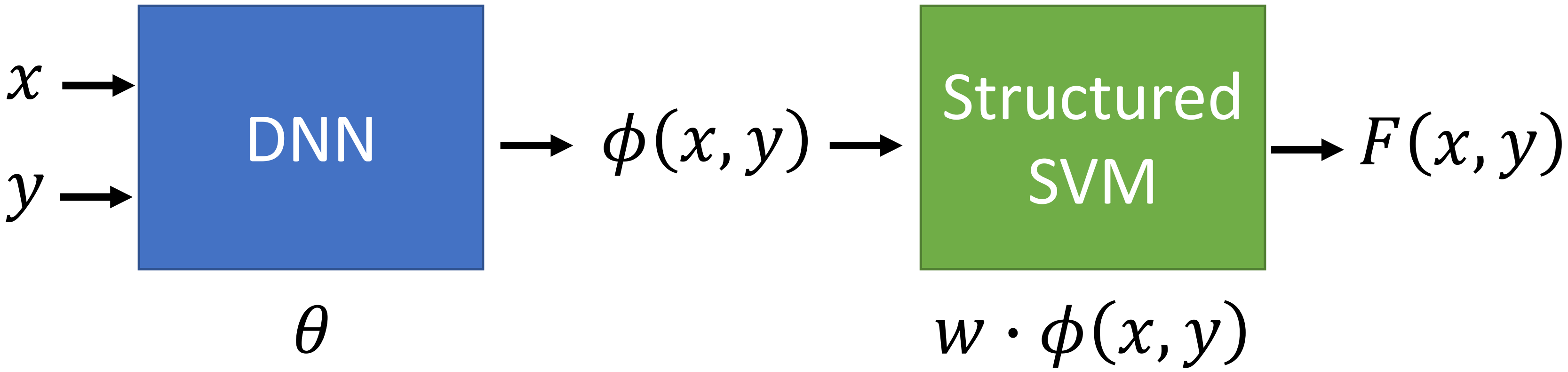
Cutting Plane Algorithm for Structured SVM

Multi-class and binary SVM

Beyond Structured SVM (open question)

Beyond Structured SVM

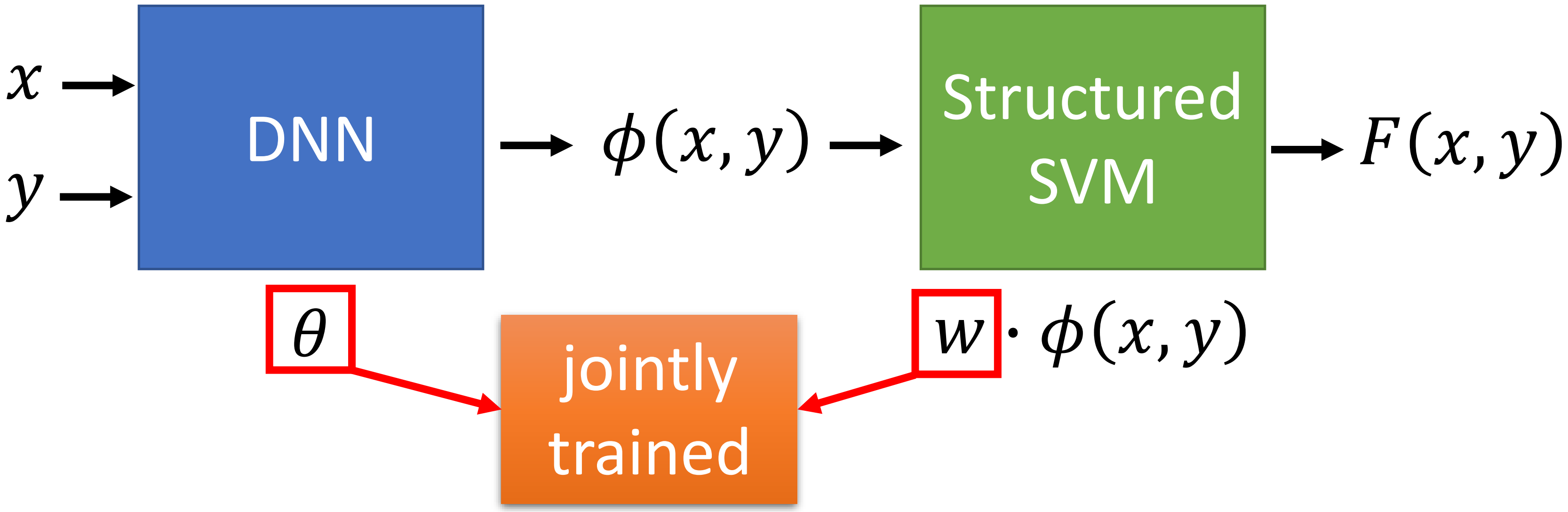
- Involving DNN when generating $\phi(x, y)$



Ref: Hao Tang, Chao-hong Meng, Lin-shan Lee, "An initial attempt for phoneme recognition using Structured Support Vector Machine (SVM)," ICASSP, 2010
Shi-Xiong Zhang, Gales, M.J.F., "Structured SVMs for Automatic Speech Recognition," in Audio, Speech, and Language Processing, IEEE Transactions on, vol.21, no.3, pp.544-555, March 2013

Beyond Structured SVM

- Jointly training structured SVM and DNN

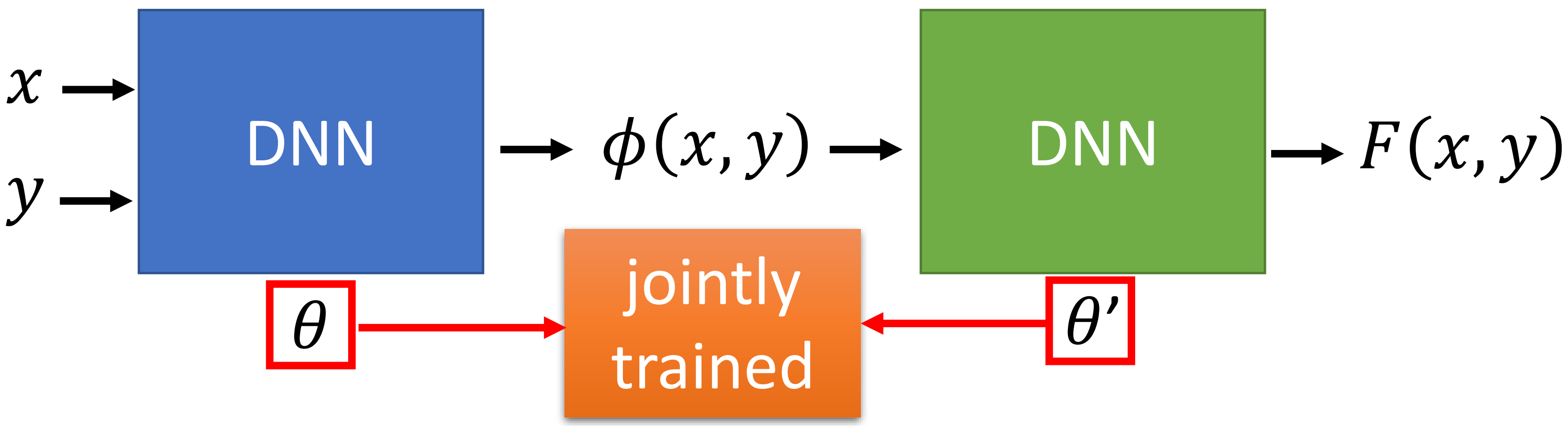


Ref: Shi-Xiong Zhang, Chaojun Liu, Kaisheng Yao, and Yifan Gong, "DEEP NEURAL SUPPORT VECTOR MACHINES FOR SPEECH RECOGNITION", Interspeech 2015

Beyond Structured SVM

- Replacing Structured SVM with DNN

A DNN with x and y as input and $F(x, y)$ (a scalar) as output



$$C = \frac{1}{2}||\theta||^2 + \frac{1}{2}||\theta'||^2 + \lambda \sum_{n=1}^{N} C^n$$

$$C^n = \max_y [\Delta(\hat{y}^n, y) + F(x^n, y)] - F(x^n, \hat{y}^n)$$

Ref: Yi-Hsiu Liao, Hung-yi Lee, Lin-shan Lee, "Towards Structured Deep Neural Network for Automatic Speech Recognition", ASRU 2015

http://speech.ee.ntu.edu.tw/~tlkagk/paper/DNN_ASRU15.pdf

Concluding Remarks

Separable case

Non-separable case

Considering Errors

Regularization

Structured SVM

Cutting Plane Algorithm for Structured SVM

Multi-class and binary SVM

Beyond Structured SVM (open question)

Acknowledgement

- 感謝 盧柏儒 同學於上課時發現投影片上的錯誤
- 感謝 徐翊祥 同學於上課時發現投影片上的錯誤