Two SVM tutorials linked in class website (please, read both):

- High-level presentation with applications (Hearst 1998)
- Detailed tutorial (Burges 1998)

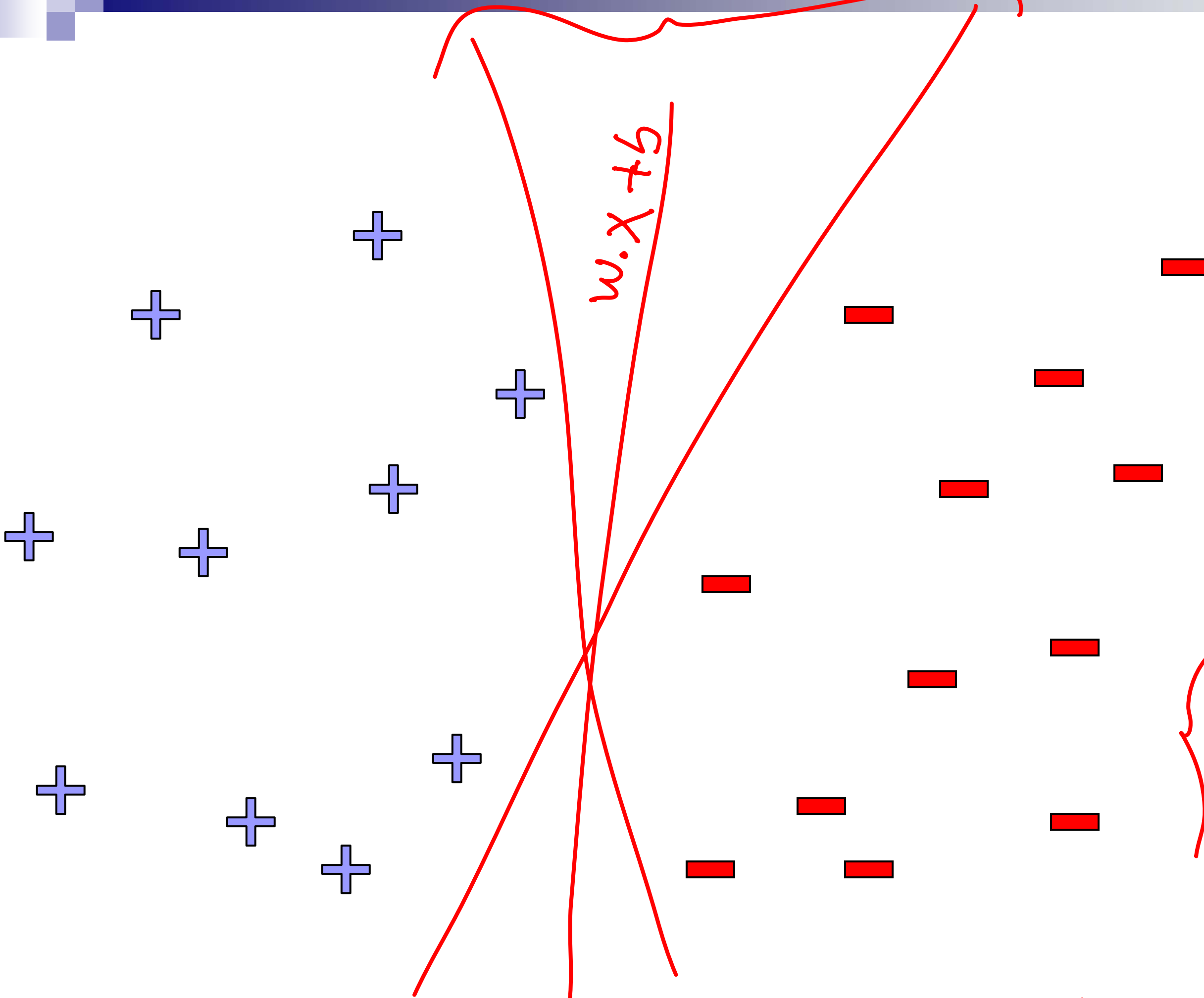
Support Vector Machines

Machine Learning – 10701/15781
Carlos Guestrin
Carnegie Mellon University

February 16th, 2005

Linear classifiers – Which line is better?

better?



Data:

$$\left\{ \begin{array}{c} \left\langle x_1^{(1)}, \ldots, x_1^{(m)}, y_1 \right\rangle \\ \vdots \\ \left\langle x_n^{(1)}, \ldots, x_n^{(m)}, y_n \right\rangle \end{array} \right.$$

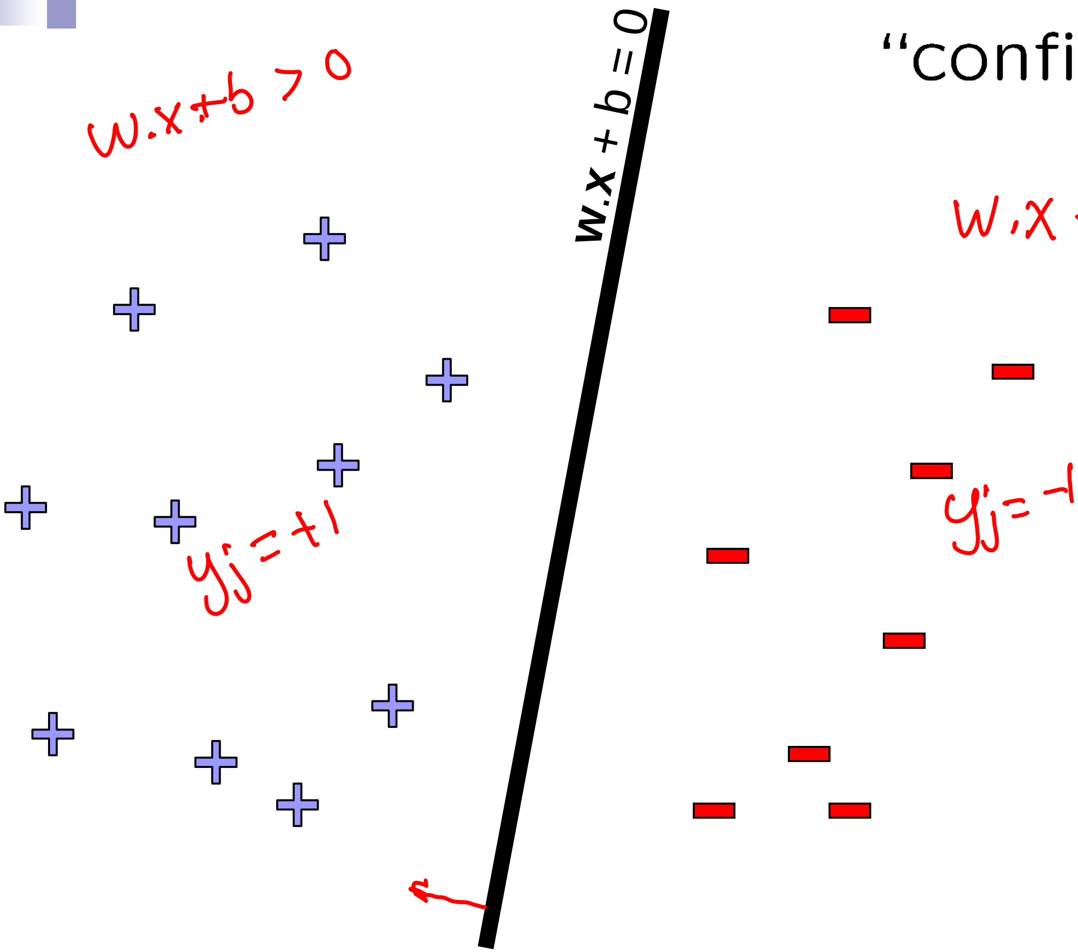
Example i:

$\left\langle x_i^{(1)}, \ldots, x_i^{(m)} \right\rangle$ — m features

$y_i \in \{-1, +1\}$ — class

w.x = $\sum_j w^{(j)} x^{(j)}$ ← dot product

Pick the one with the largest margin!



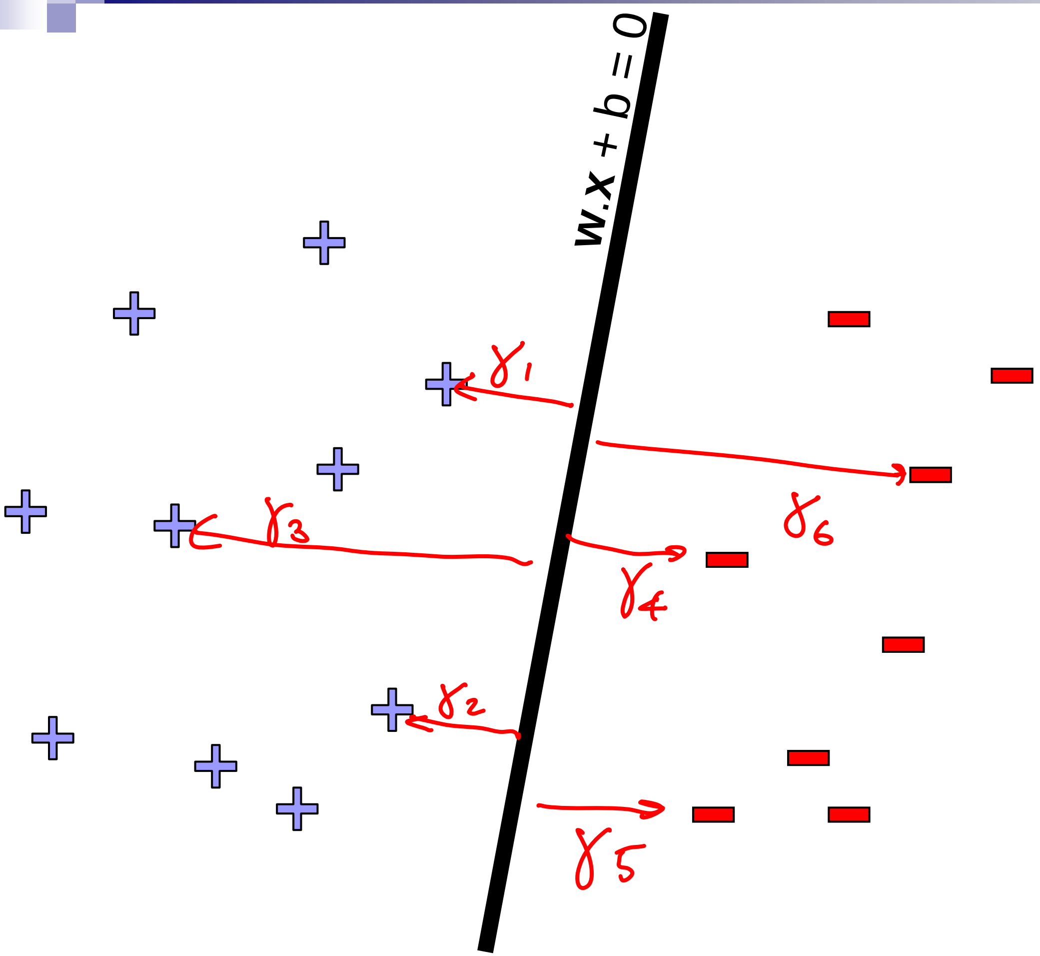
"confidence" $= \underbrace{\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j}$

confidence $<x_j, y_j>$

$(w.x_j+b).y_j$ is big

$\mathbf{w.x} = \sum_j w^{(j)} x^{(j)}$

Maximize the margin



$$(x_1 . w + b)\, y_1 = \gamma_1$$

$$(x_j . w + b) . y_j = \gamma_j$$

$$\vdots$$

$$(x_n . w + b)\, y_n = \gamma_n$$

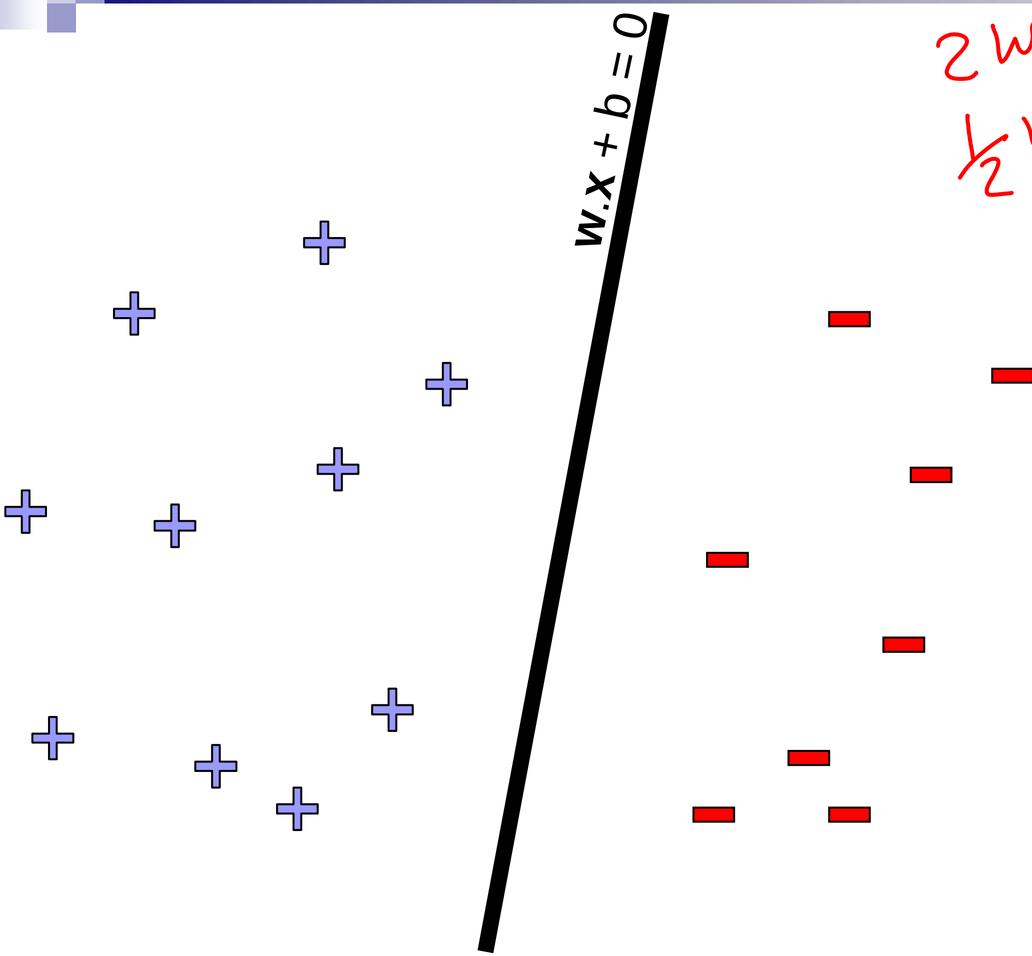
$$\gamma = \min_j \gamma_j$$

$$\boxed{\underset{w,b}{\text{maximize}}\ \gamma}$$

$$(x_j\, w + b)\, y_j \geq \gamma \quad \forall j$$

But there are a many planes…

maximize γ



$2wx + 2b = 0$

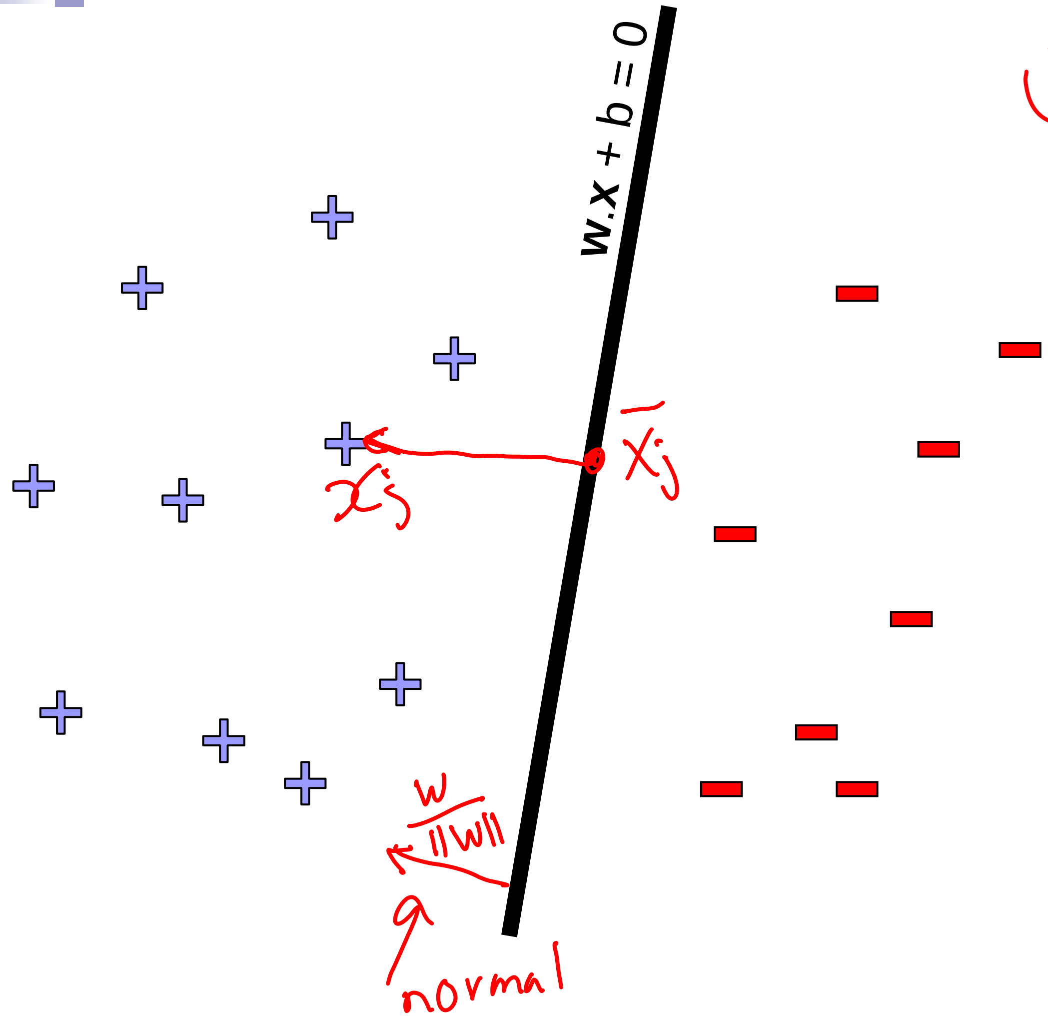
$\frac{1}{2} w.x + \frac{1}{2} b = 0$

Redundancy

$\max_{w,b} \ \gamma$

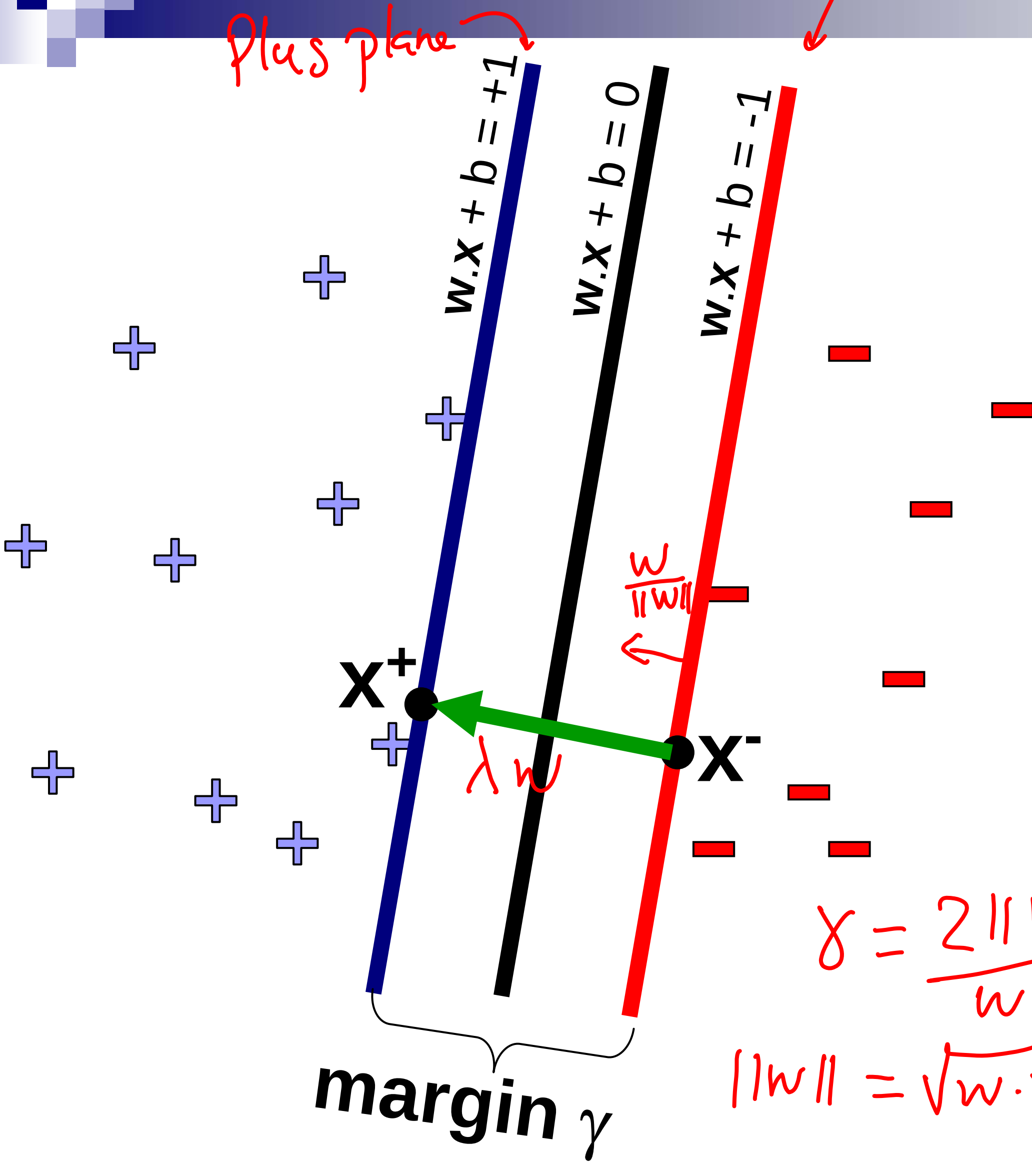
$(x_j w + b) y_j \geq \gamma \ \forall j$

Review: Normal to a plane



$$\underbrace{x_j = \bar{x}_j + \lambda . w}_{\text{any point in the space}}$$

Normalized margin – Canonical hyperplanes



$x^+ = x^- + \lambda w$

↳ Plug into Plus plane:

$w x^+ + b = +1$

$w(x^- + \lambda w) + b = +1$

$\underbrace{w x^- + b}_{} + \lambda\, w.w = +1$

x^- in Minus Plane

-1

$-1 + \lambda\, w.w = +1$

$\lambda = \frac{2}{w.w}$

$\gamma = \frac{2\|w\|}{w.w}$

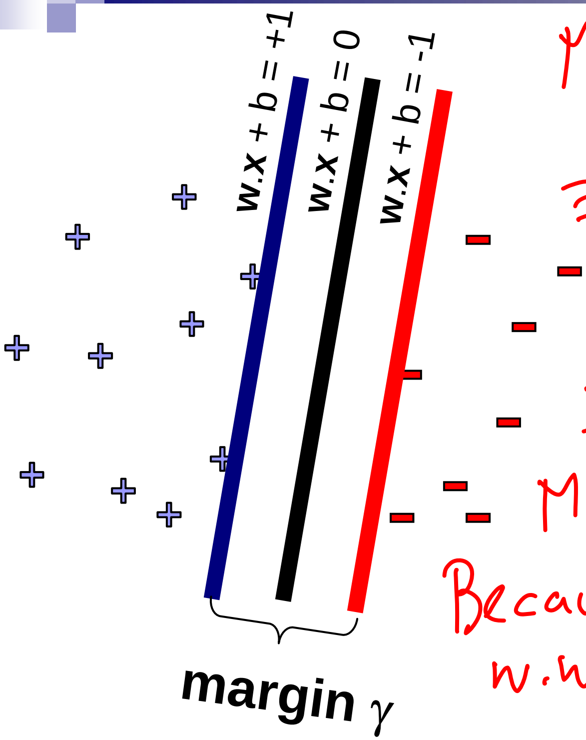
$\|w\| = \sqrt{w.w}$

margin of conf.

Max: $\boxed{\gamma = \frac{2}{\sqrt{\mathbf{w}.\mathbf{w}}}}$

Margin maximization using canonical hyperplanes

$\frac{2}{\sqrt{w.w}} \le 1$



Max $\frac{2}{\sqrt{w.w}}$

$\equiv$ Min $\frac{\sqrt{w.w}}{2}$

$\equiv$ Min $w.w$

Because $w.w \geq 0$

$\sqrt{\cdot}$ is monotonic

$\text{Maximize}_{w,b} \quad \gamma$

$(w.x_j + b) y_j \geq \gamma \quad \forall j$

subject $\gamma \le 1$

Might as well $\gamma = 1$

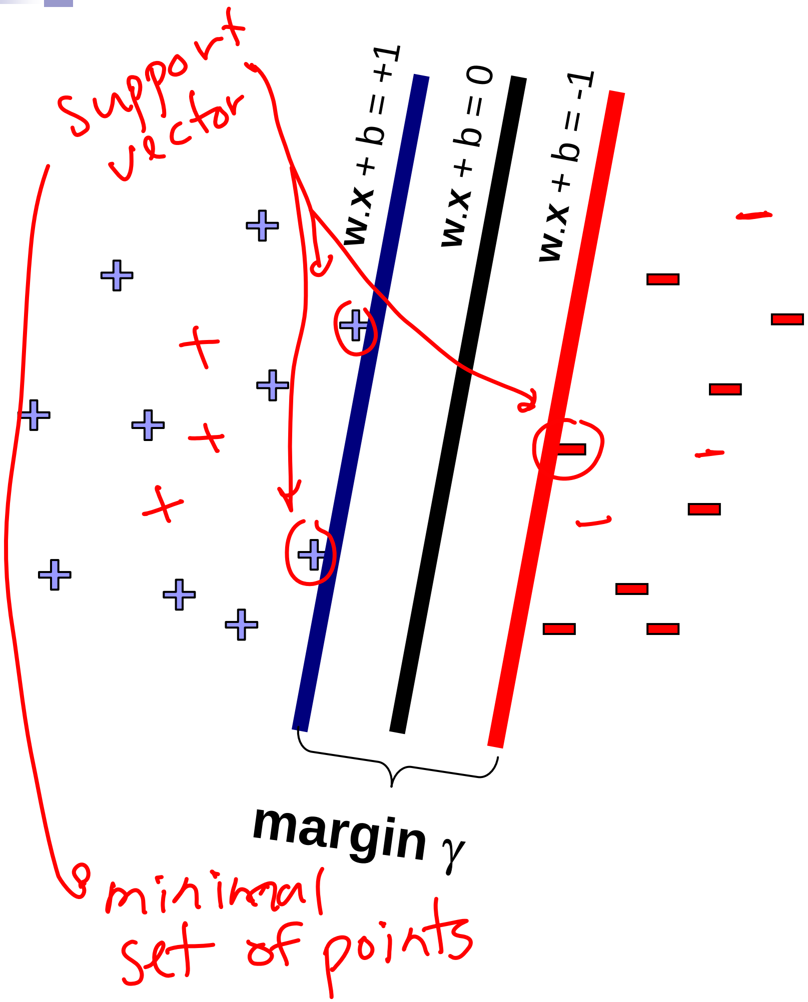
$\text{Maximize}_{w,b} \quad \frac{2}{\sqrt{w.w}}$

$(w.x_j + b) y_j \geq 1 \quad \forall j$

$$\text{minimize}_{\mathbf{w}} \quad \mathbf{w}.\mathbf{w}$$

$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j \in \text{Dataset}$$

Support vector machines (SVMs)

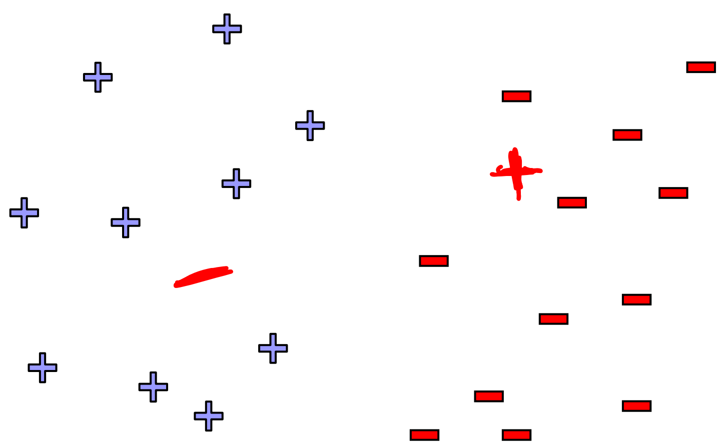


$$\text{minimize}_{\mathbf{w}} \quad \mathbf{w}.\mathbf{w}$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1, \ \forall j$$

- Solve efficiently by quadratic programming (QP)
  - Well-studied solution algorithms
- Hyperplane defined by support vectors

What if the data is not linearly separable?

Use features of features of features of features….



not linearly seperable
$\Rightarrow \nexists$ hyperplane
$\gamma > 0$

2d $\langle x^{(1)}, x^{(2)}, y \rangle$

Feed SVM:

$\langle x^{(1)}, x^{(1)} \cdot x^{(1)}, x^{(2)}, x^{(2)} \cdot x^{(2)}, y \rangle$

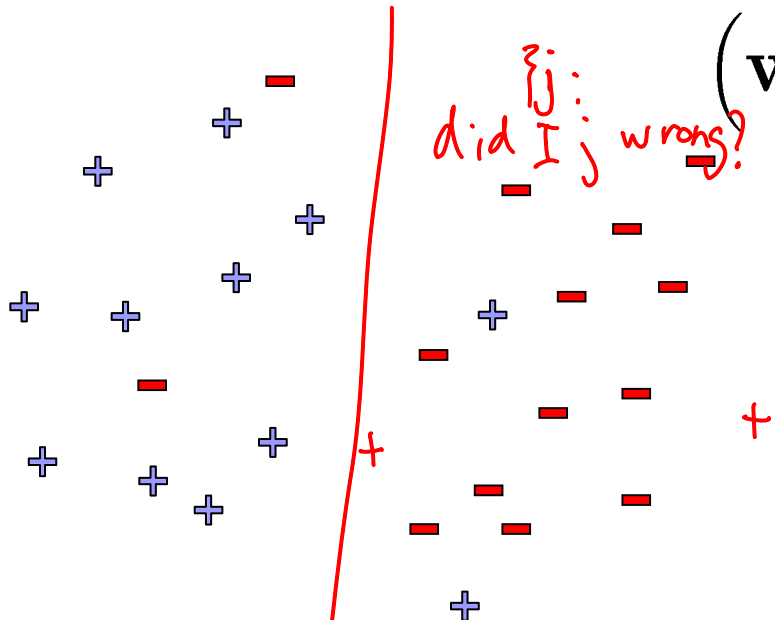
polynomial features

What if the data is still not linearly separable?

$$\text{minimize}_{\mathbf{w}} \quad \mathbf{w}.\mathbf{w} + C\,(\#\text{mistakes})$$

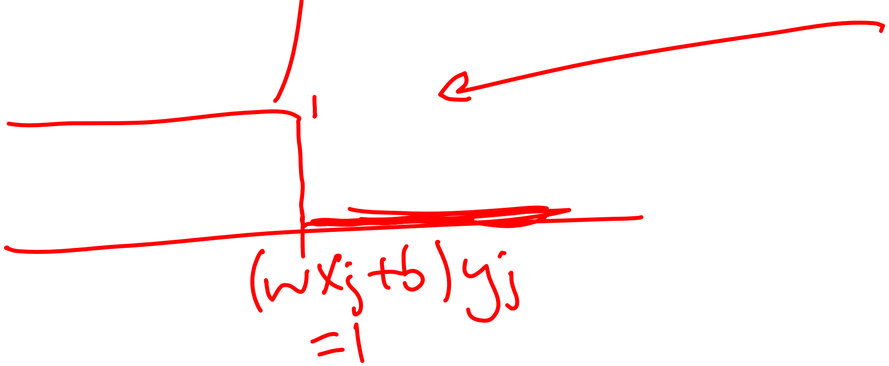
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 \qquad , \forall j$$

{j : did I j wrong?}



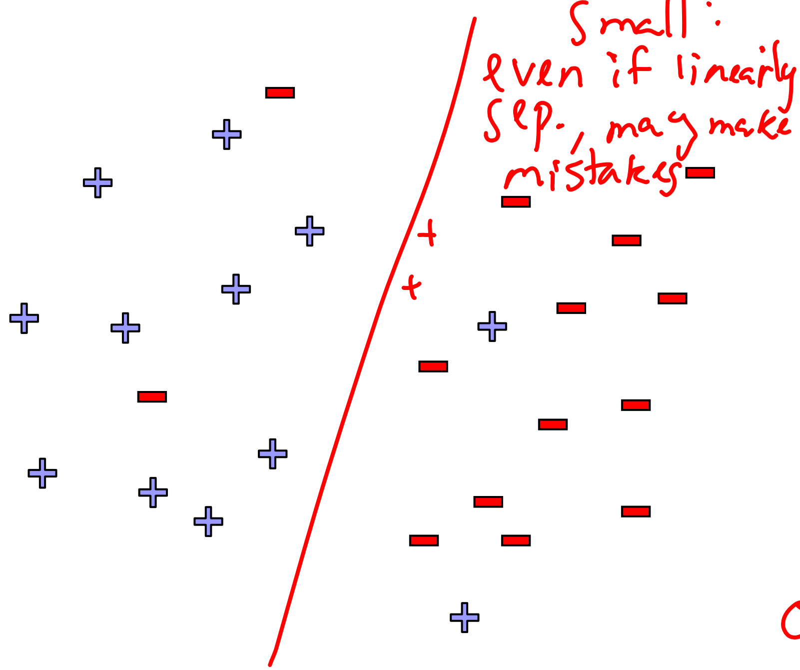
- Minimize **w.w** and number of training mistakes
  - Tradeoff two criteria?

- Tradeoff #(mistakes) and **w.w**
  - 0/1 loss
  - Slack penalty *C*
  - Not QP anymore
  - Also doesn't distinguish near misses and really bad mistakes



Slack variables – Hinge loss

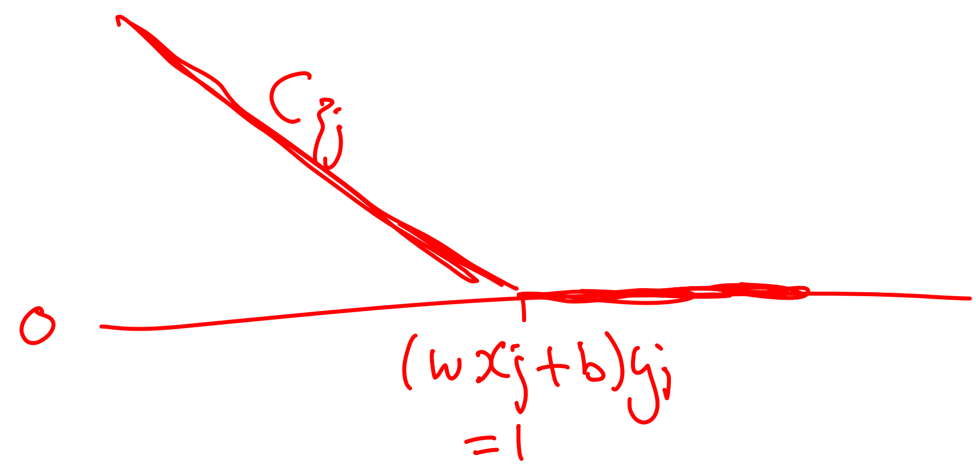
if C small: even if linearly sep., may make mistakes



$$\text{minimize}_{\mathbf{w}} \quad \mathbf{w}.\mathbf{w} + C \sum_j \xi_j$$

$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j \quad, \forall j$$

$$\xi_j \geq 0$$



- If margin $\geq$ 1, don't care $\Rightarrow \xi_j = 0$, pay nothing
- If margin < 1, pay linear penalty $\Rightarrow \xi_j > 0$, and pay $C \cdot \xi_j$

Side note: What's the difference between SVMs and logistic regression?

SVM: linear classifiers **Logistic regression:**

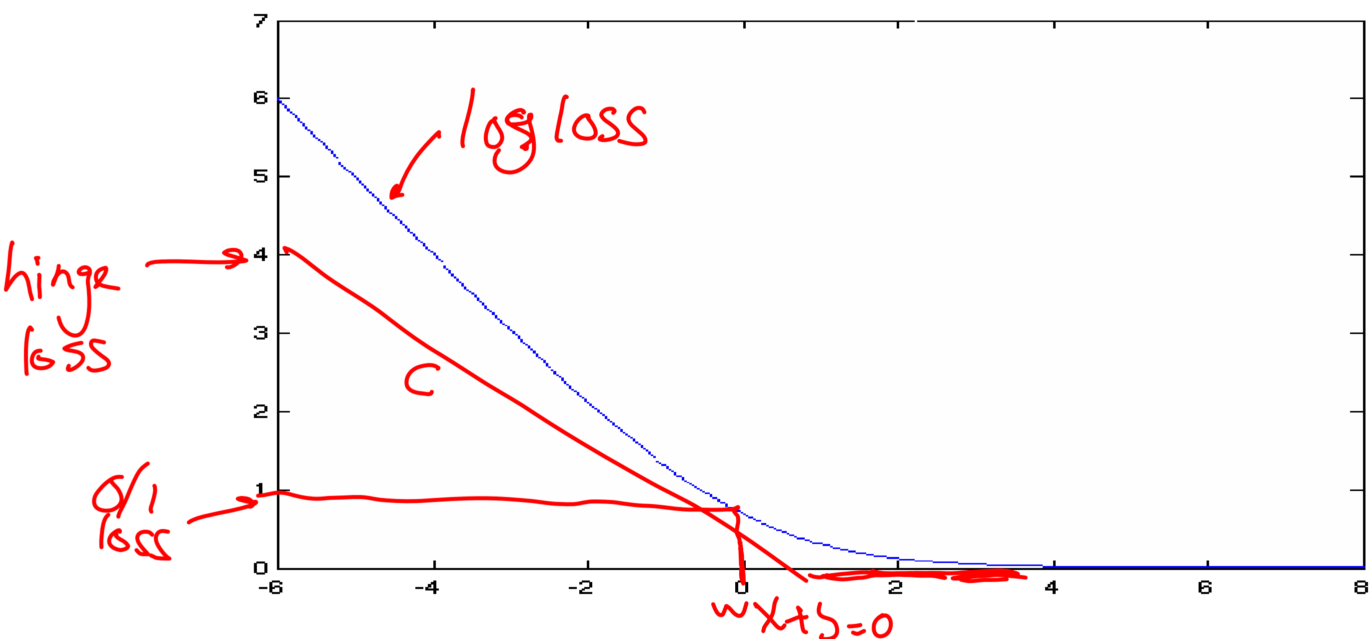
$$\text{minimize}_{\mathbf{w}} \quad \mathbf{w}.\mathbf{w} + C\sum_j \xi_j$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j, \ \ \forall j$$
$$\xi_j \geq 0, \ \ \forall j$$

Hinge Loss:

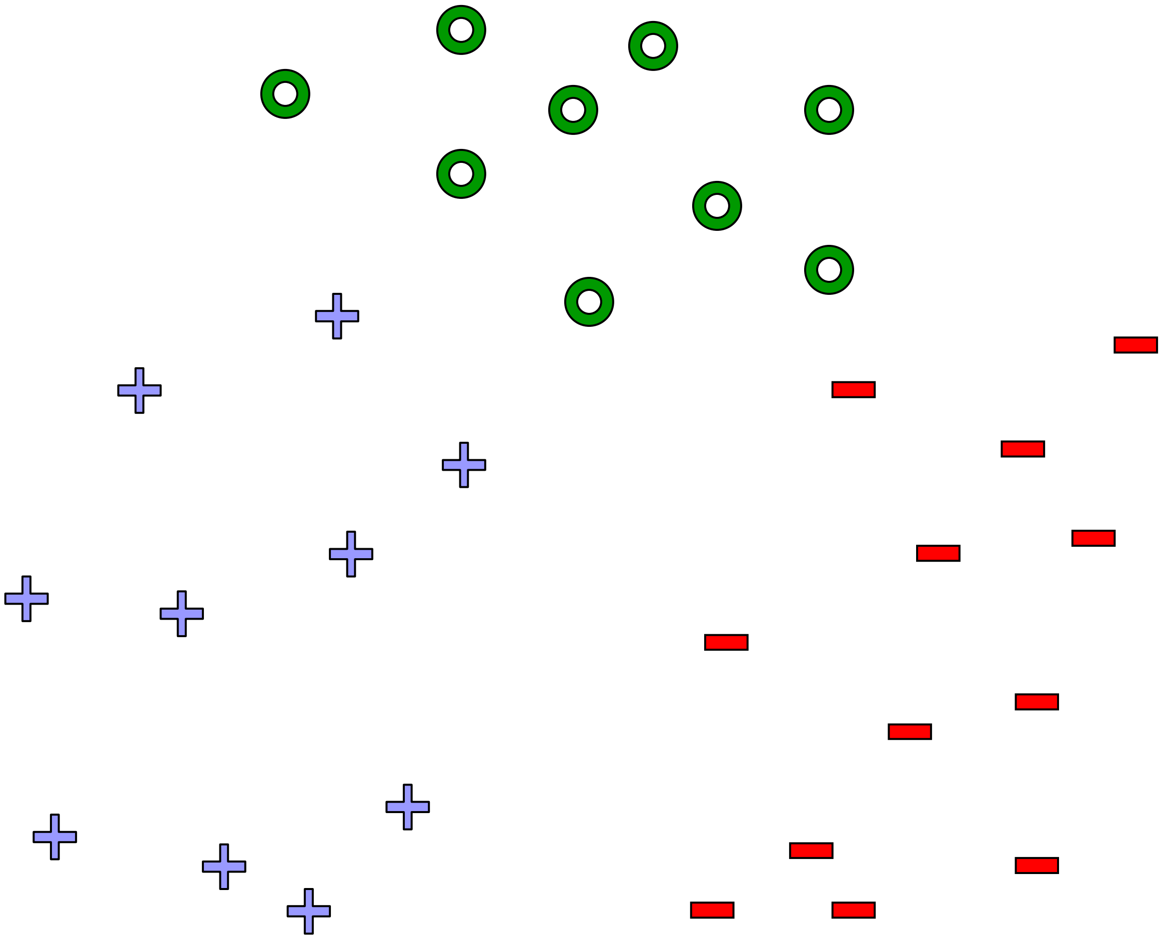
$$P(Y=1 \mid x, \mathbf{w}) \ = \ \frac{1}{1+e^{-(\mathbf{w}.\mathbf{x}+b)}}$$

Log loss:

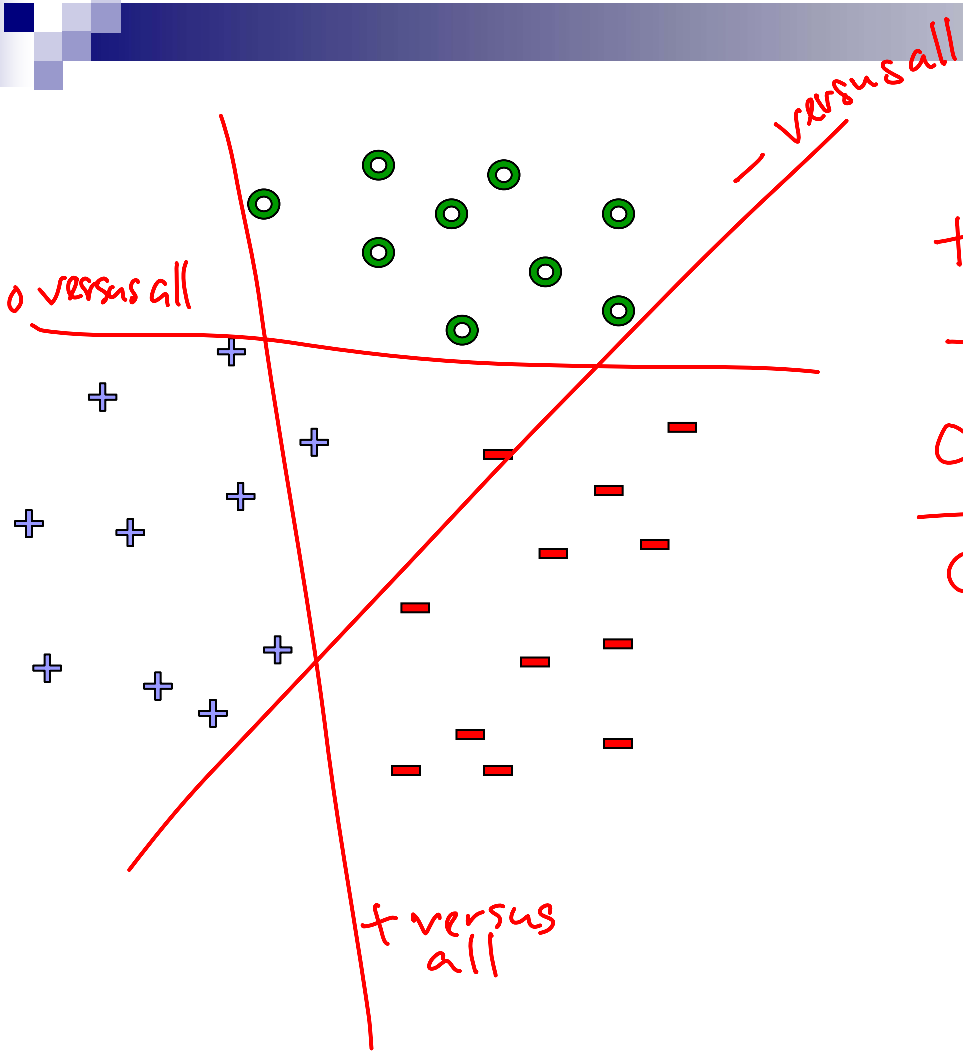
$$\min \ -\ln P(Y=1 \mid x, \mathbf{w}) \ = \ \ln\left(1+e^{-(\mathbf{w}.\mathbf{x}+b)}\right)$$



What about multiple classes?



One against All



Learn 3 classifiers:

+ versus {o, −}

− versus {o, +}

o versus {−, +}

---

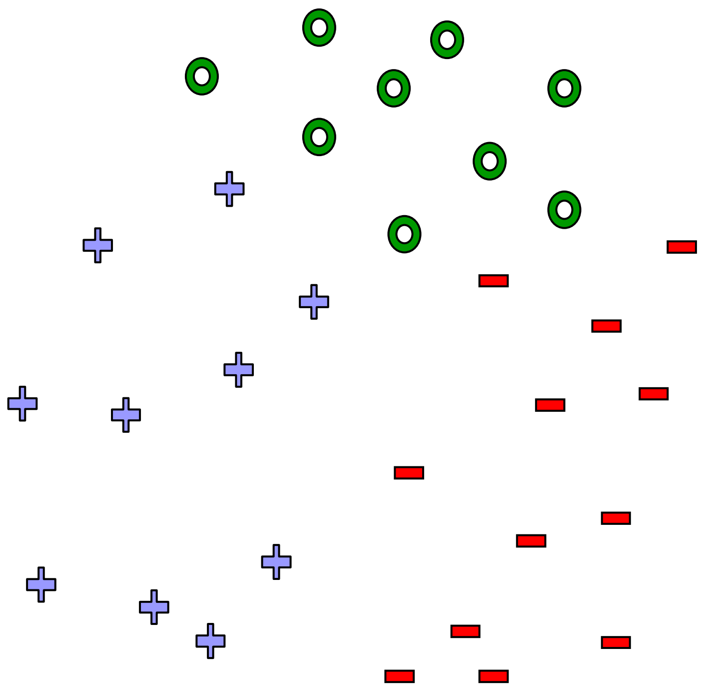
class x

classifier with highest confidence

Learn 1 classifier: Multiclass SVM

$$\left\{\begin{array}{l} w^{(+)}, b^{(+)} \\ w^{(0)}, b^{(0)} \\ w^{(-)}, b^{(-)} \end{array}\right.$$

Simultaneously learn 3 sets of weights



For + examples:

$$(w^{(+)} x_j + b^{(+)}) \geq 1 + w^{(-)} x_j + b^{(-)}$$

$$w^{(+)} x_j + b^{(+)} \geq 1 + w^{(0)} x_j + b^{(0)}$$

For − examples

$w^{(-)}, b^{(-)}$ win

For 0 examples

$w^{(0)}, b^{(0)}$ win

$$\boxed{\mathbf{w}^{(y_j)}.\mathbf{x}_j + b^{(y_j)} \geq \mathbf{w}^{(y')}.\mathbf{x}_j + b^{(y')} + 1, \ \forall y' \neq y_j, \ \forall j}$$

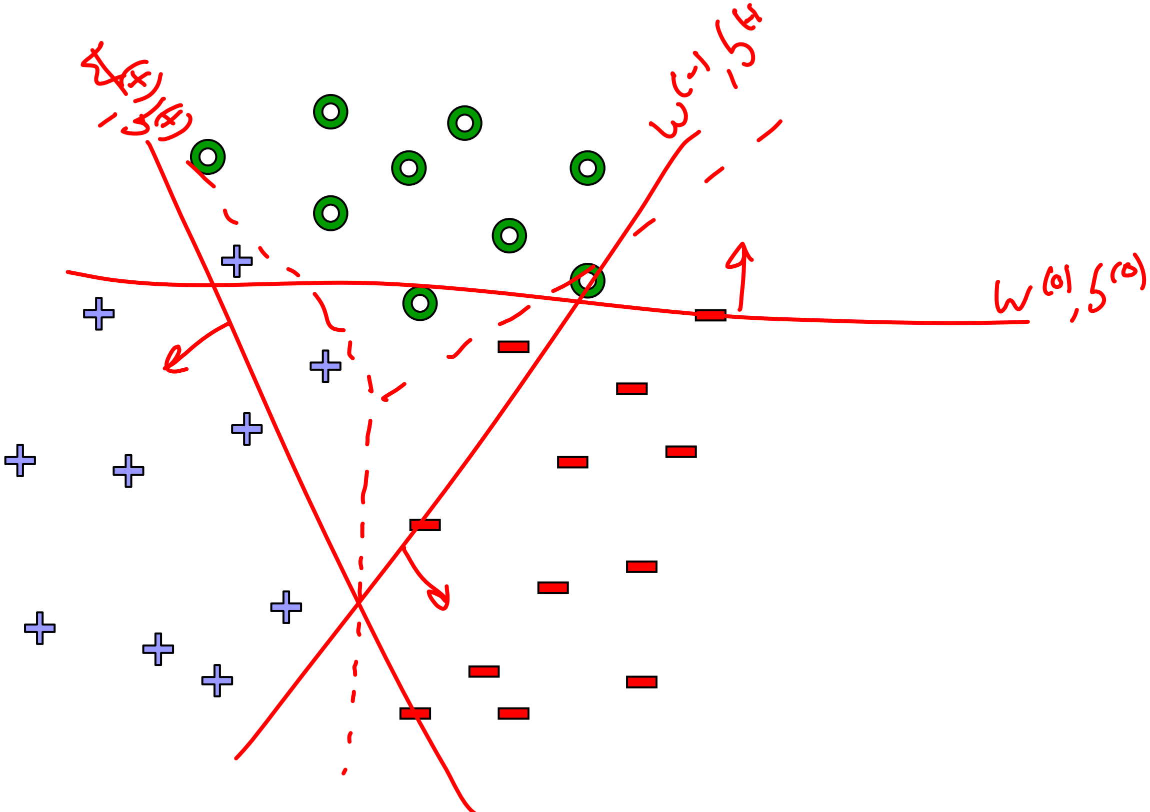
Learn 1 classifier: Multiclass SVM

possible classes

$$\text{minimize}_{\mathbf{w}} \quad \sum_y \mathbf{w}^{(y)}.\mathbf{w}^{(y)} + C\sum_j \xi_j$$

$$\mathbf{w}^{(y_j)}.\mathbf{x}_j + b^{(y_j)} \geq \mathbf{w}^{(y')}.\mathbf{x}_j + b^{(y')} + 1 - \xi_j,\ \forall y' \neq y_j,\ \forall j$$

$$\xi_j \geq 0,\ \forall j$$



What you need to know

- Maximizing margin
- Derivation of SVM formulation
- Slack variables and hinge loss
- Relationship between SVMs and logistic regression
  - 0/1 loss
  - Hinge loss
  - Log loss
- Tackling multiple class
  - One against All
  - Multiclass SVMs

Acknowledgment

- SVM applet:
  - http://www.site.uottawa.ca/~gcaron/applets.htm