Support Vector Machines , SVMs

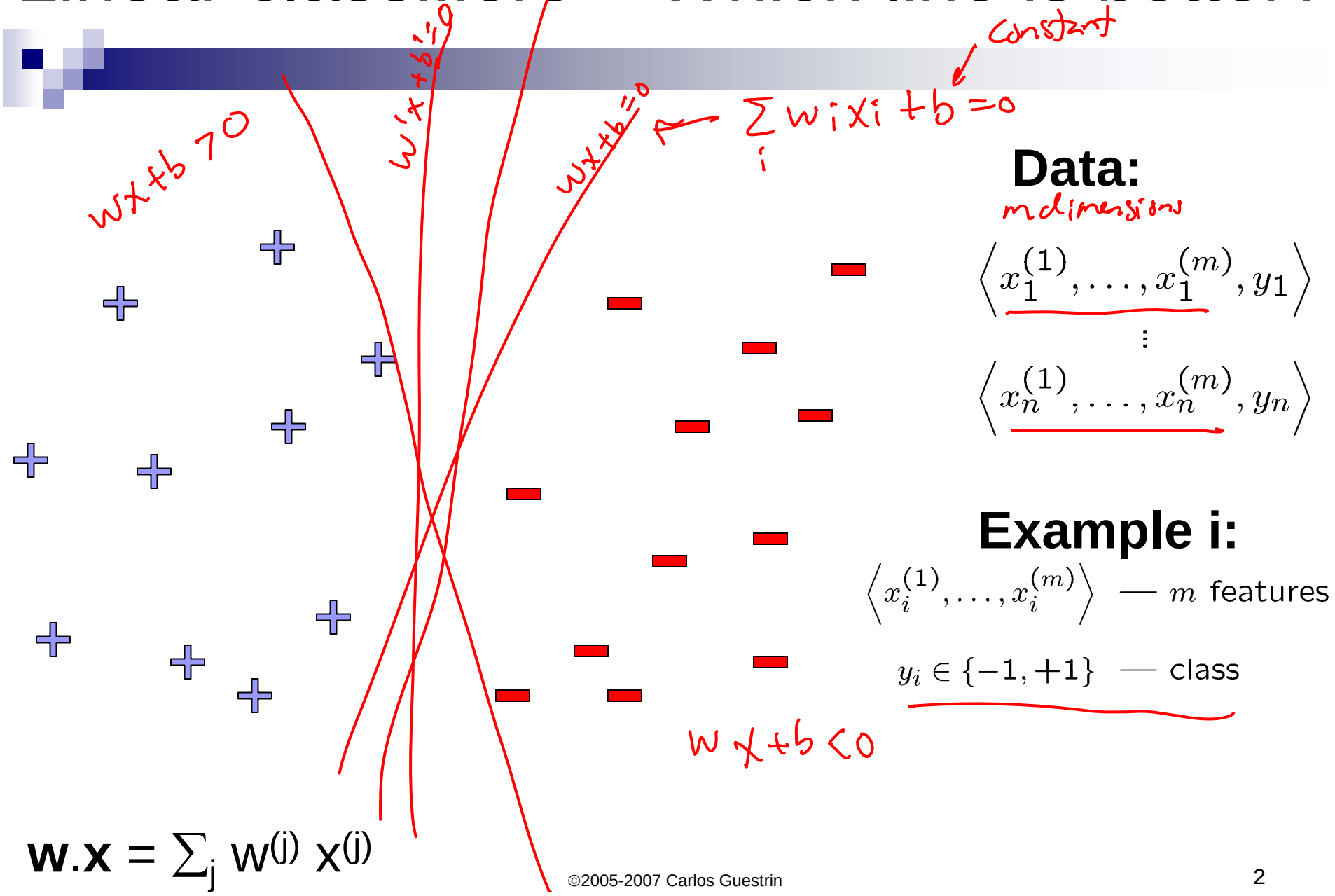
Machine Learning – 10701/15781

Carlos Guestrin

Carnegie Mellon University

February 21st, 2007

Linear classifiers – Which line is better?



Data:

m dimensions

$$\langle x_1^{(1)}, \ldots, x_1^{(m)}, y_1 \rangle$$
$$\vdots$$
$$\langle x_n^{(1)}, \ldots, x_n^{(m)}, y_n \rangle$$

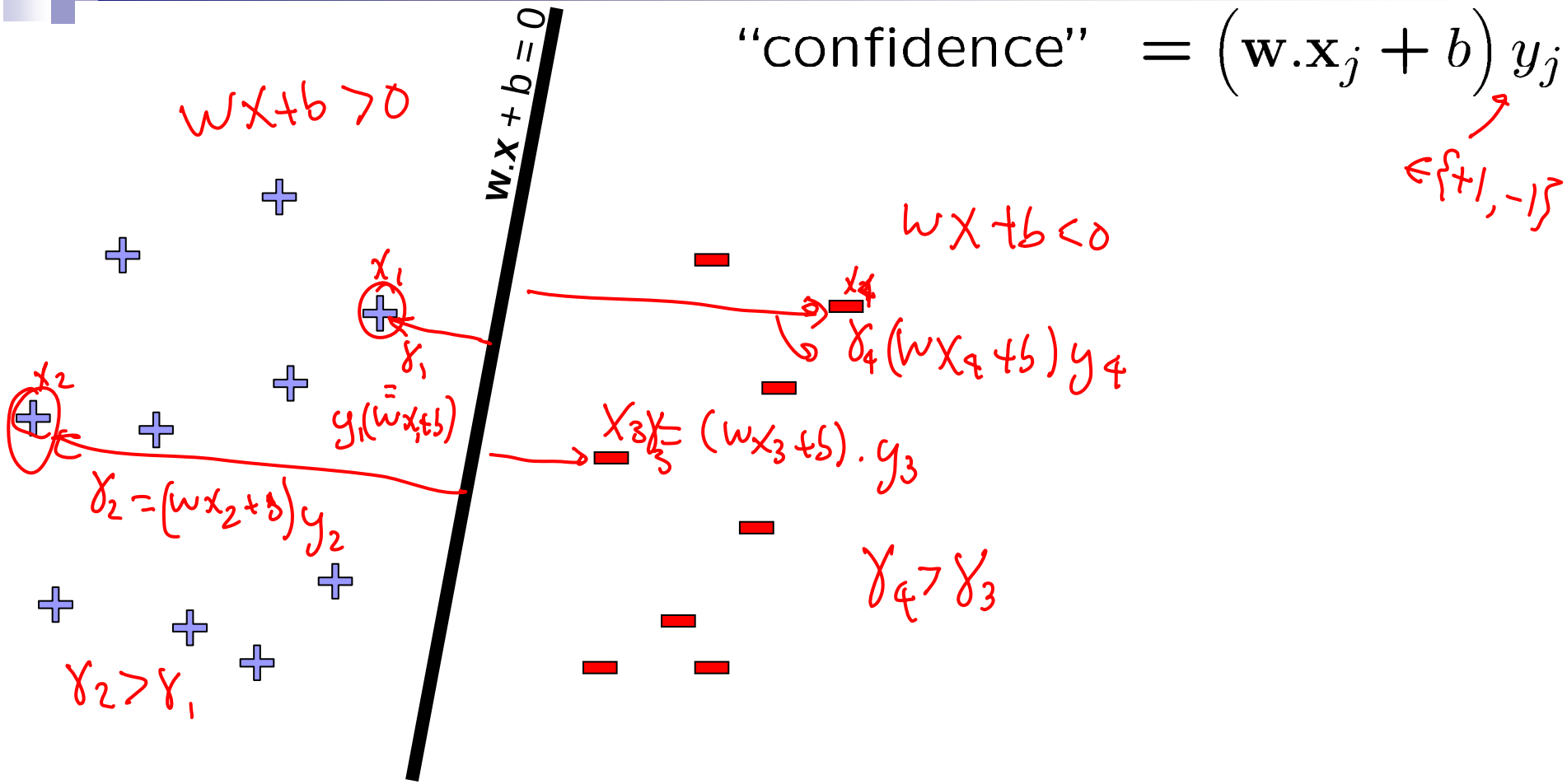
Example i:

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

$y_i \in \{-1, +1\}$ — class

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

Pick the one with the largest margin!

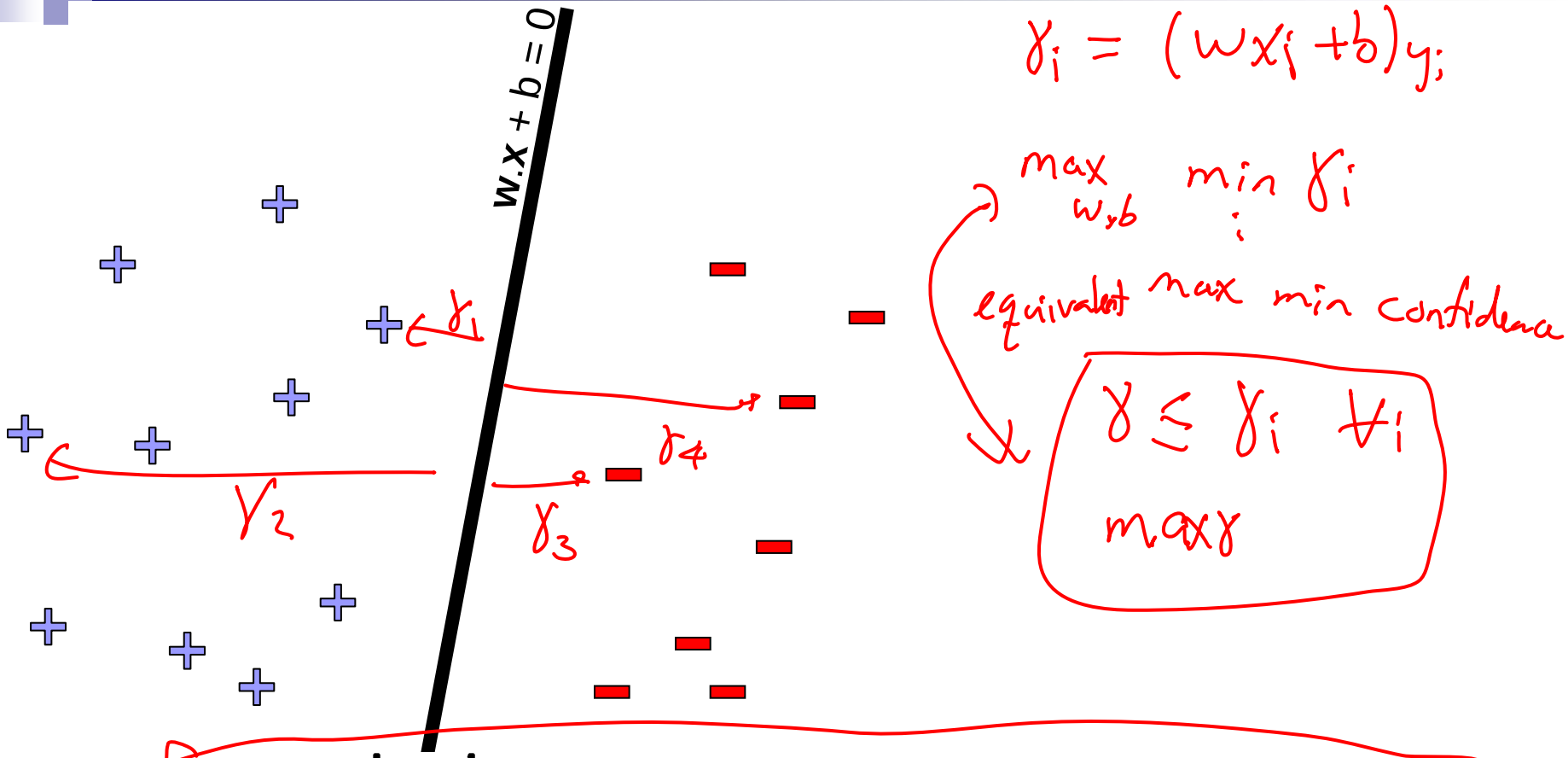


"confidence" $= \left(\mathbf{w}.\mathbf{x}_j + b\right) y_j$

$\in \{+1, -1\}$

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

Maximize the margin



$\gamma_i = (w x_i + b) y_i$

$\max_{w,b} \min \gamma_i$

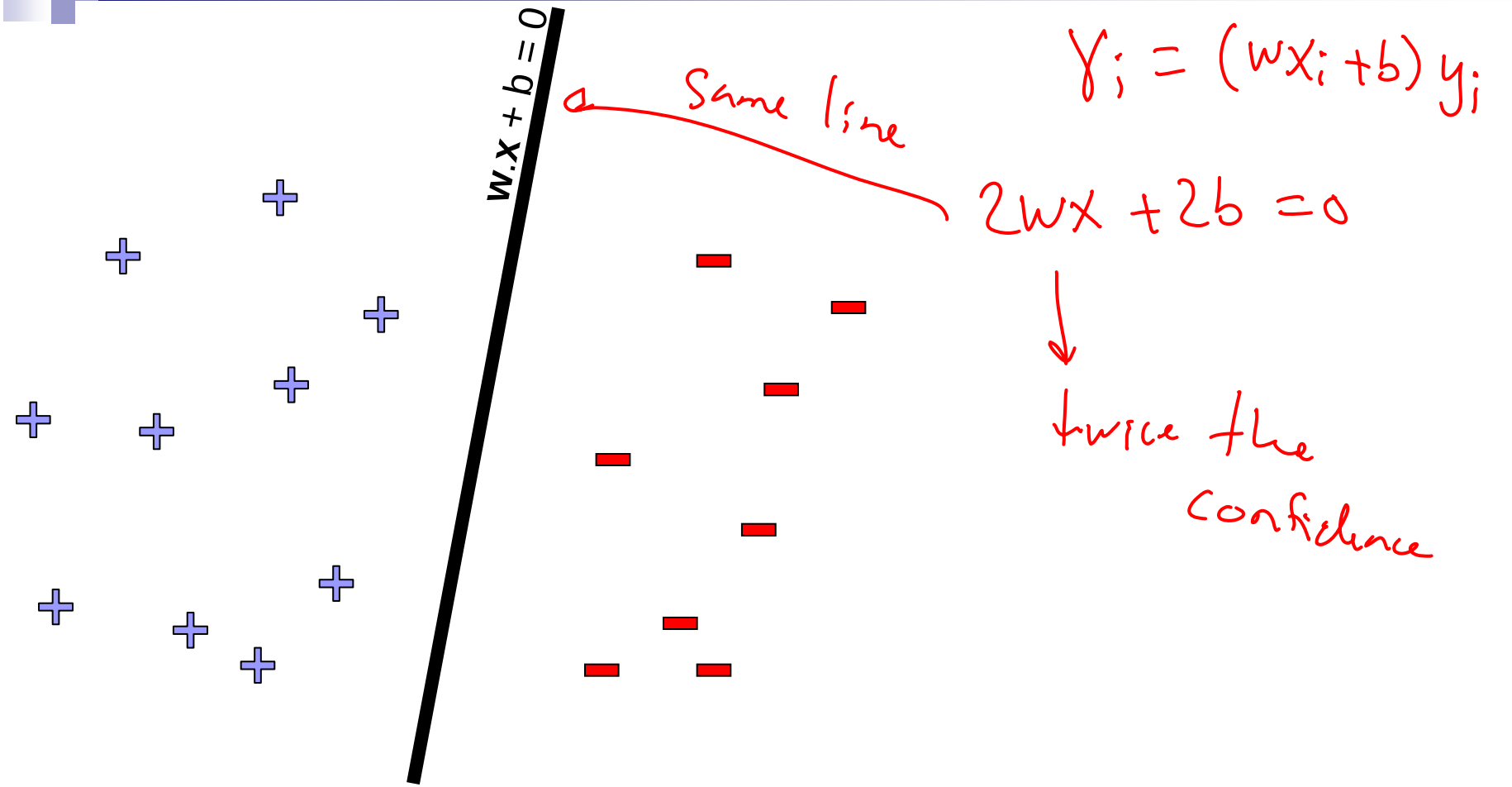
equivalent max min confidence

$\gamma \leq \gamma_i \quad \forall i$

max γ

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

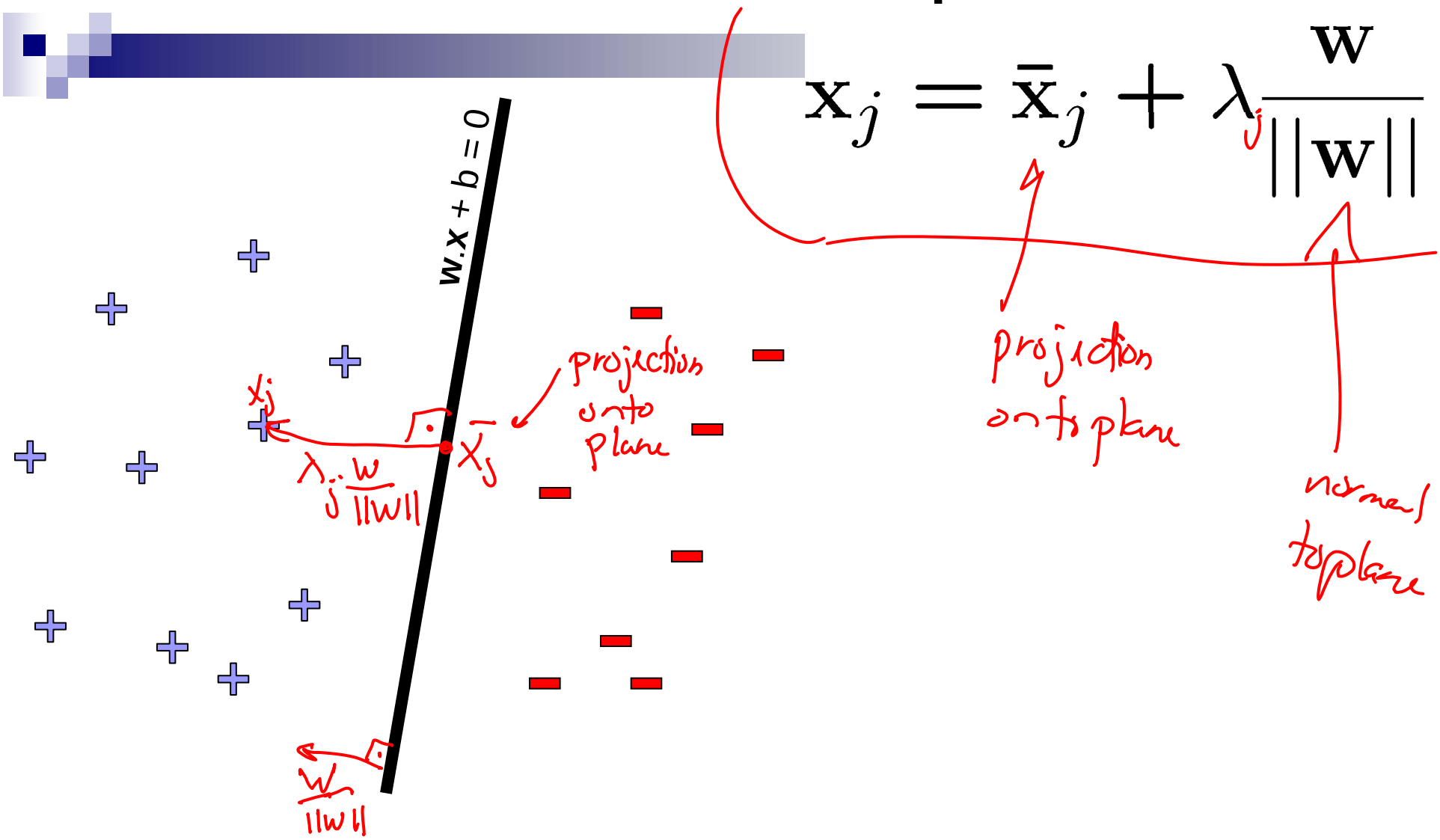
But there are a many planes…



$$\gamma_i = (wx_i + b) y_i$$

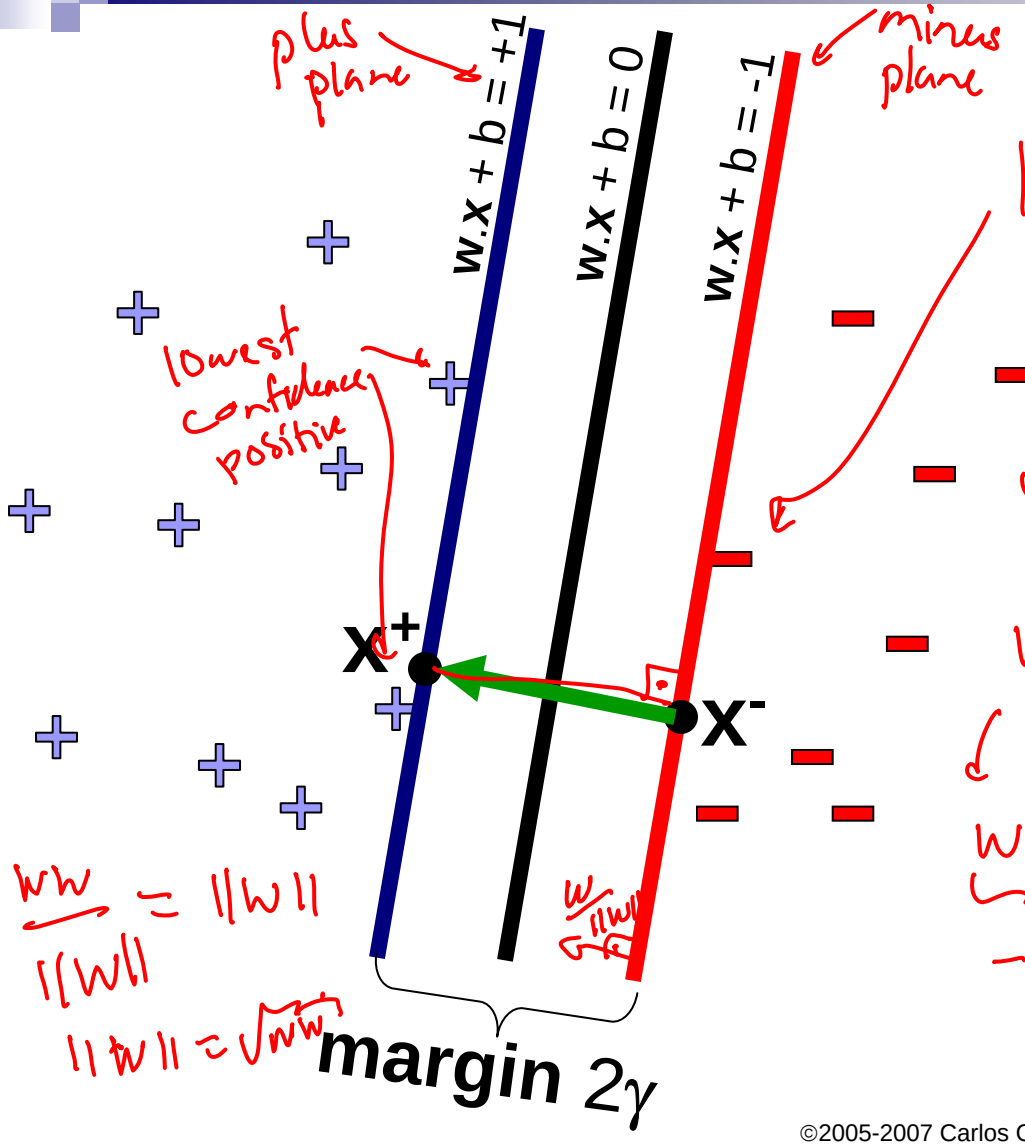
Review: Normal to a plane

$$\mathbf{x}_j = \bar{\mathbf{x}}_j + \lambda_j \frac{\mathbf{w}}{||\mathbf{w}||}$$



Normalized margin – Canonical hyperplanes

$$\mathbf{x}_j = \bar{\mathbf{x}}_j + \lambda \frac{\mathbf{w}}{||\mathbf{w}||}$$



$x^+ = x^- + \lambda \frac{w}{||w||}$

substitute

$w x^+ + b = 1$

$w\left(x^- + \lambda \frac{w}{||w||}\right) + b = 1$

$w x^- + b + \lambda \frac{ww}{||w||} = 1$

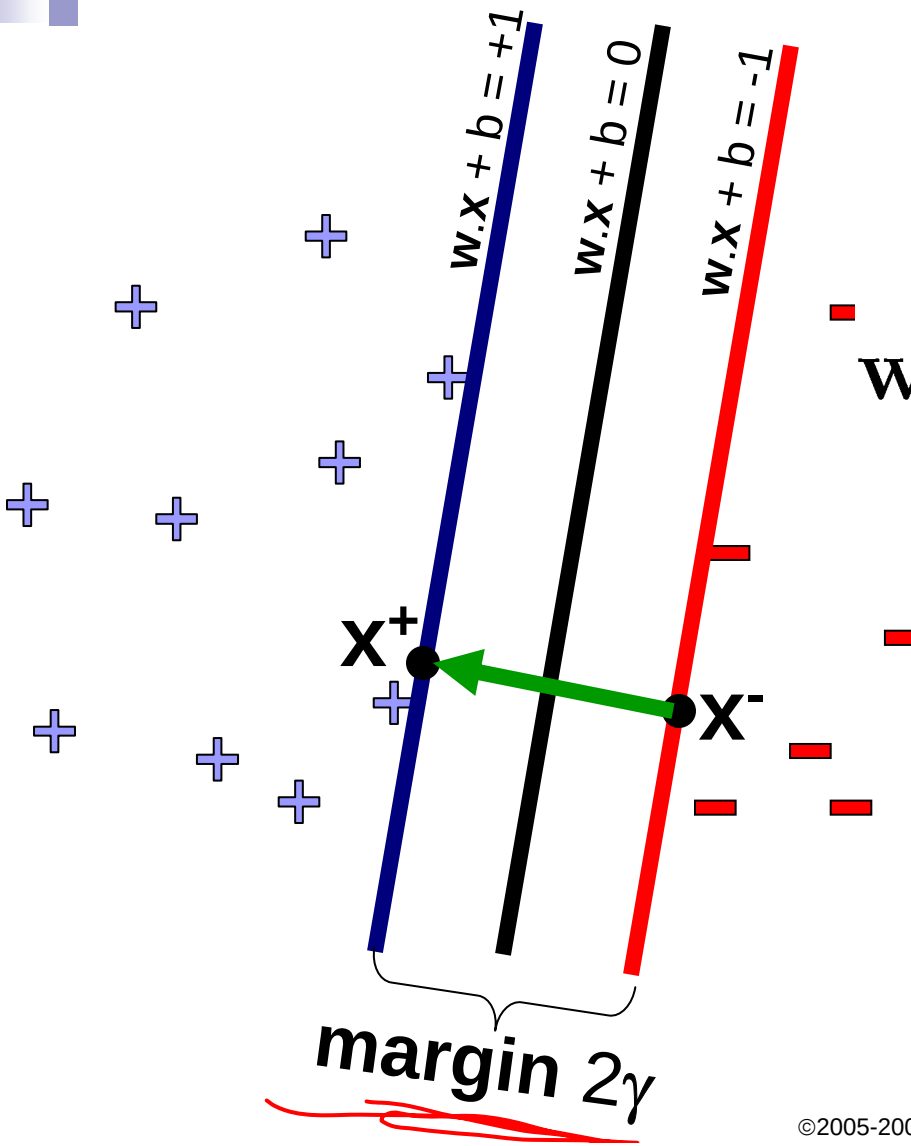
$w x^- + b = -1$

$\frac{ww}{||w||} = ||w||$

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

$\lambda = \frac{2}{ww/||w||} = \frac{2}{||w||}$

Normalized margin – Canonical hyperplanes



$$\mathbf{x}^+ = \mathbf{x}^- + \lambda \mathbf{w}$$

x^+ belongs to + plane

$$\mathbf{w}.\mathbf{x}^+ + b = 1$$

substitute

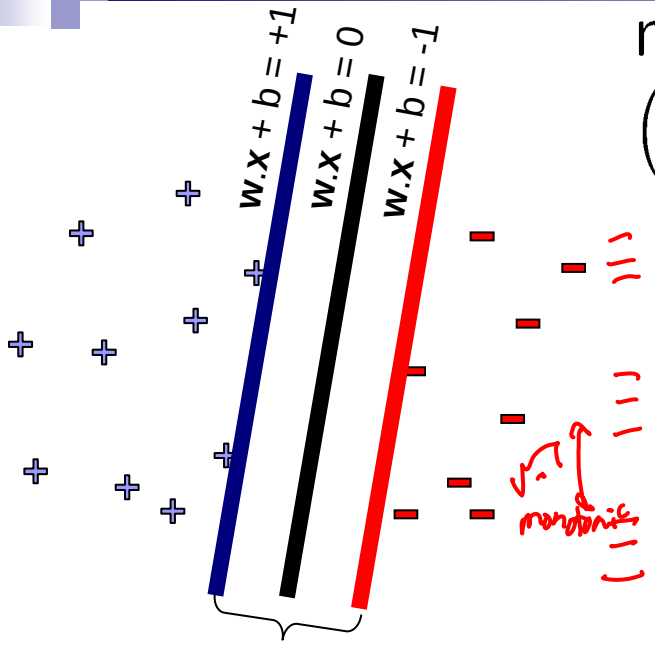
$$\mathbf{w}.(\mathbf{x}^- + \lambda \frac{\mathbf{w}}{||\mathbf{w}||}) + b = 1$$

$$\lambda = \frac{2}{||\mathbf{w}||}$$

$$\gamma = \frac{1}{\sqrt{\mathbf{w}.\mathbf{w}}} = \frac{1}{||\mathbf{w}||}$$

Margin maximization using canonical hyperplanes

$$\gamma = \frac{1}{\sqrt{\mathbf{w}.\mathbf{w}}}$$



$$\text{maximize}_{\gamma,\mathbf{w},b} \quad \gamma$$

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

$\gamma \leq 1$

$\equiv$ max $\frac{1}{\sqrt{w.w}}$

$\equiv$ min $\sqrt{w.w}$

$\equiv$ min $w.w$

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

$\gamma \leq 1$

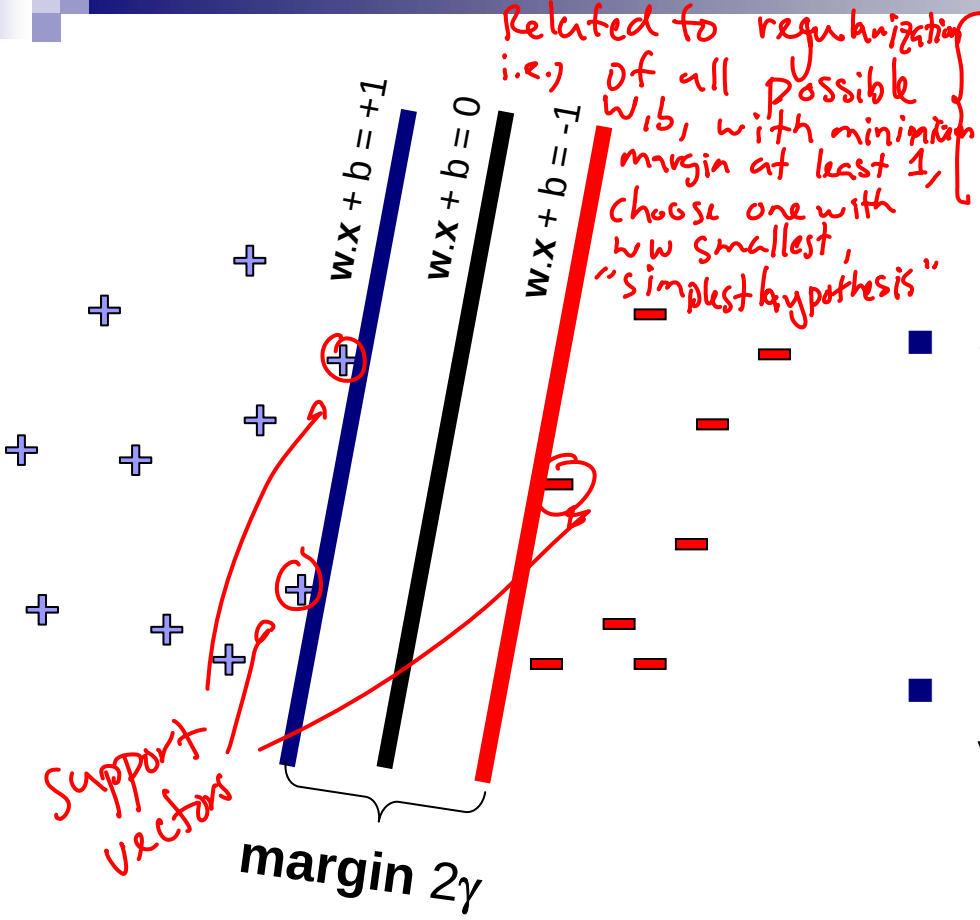
will maximize γ until hit $\gamma = 1$

Confidence of every point is at least 1.

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

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

Support vector machines (SVMs)



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

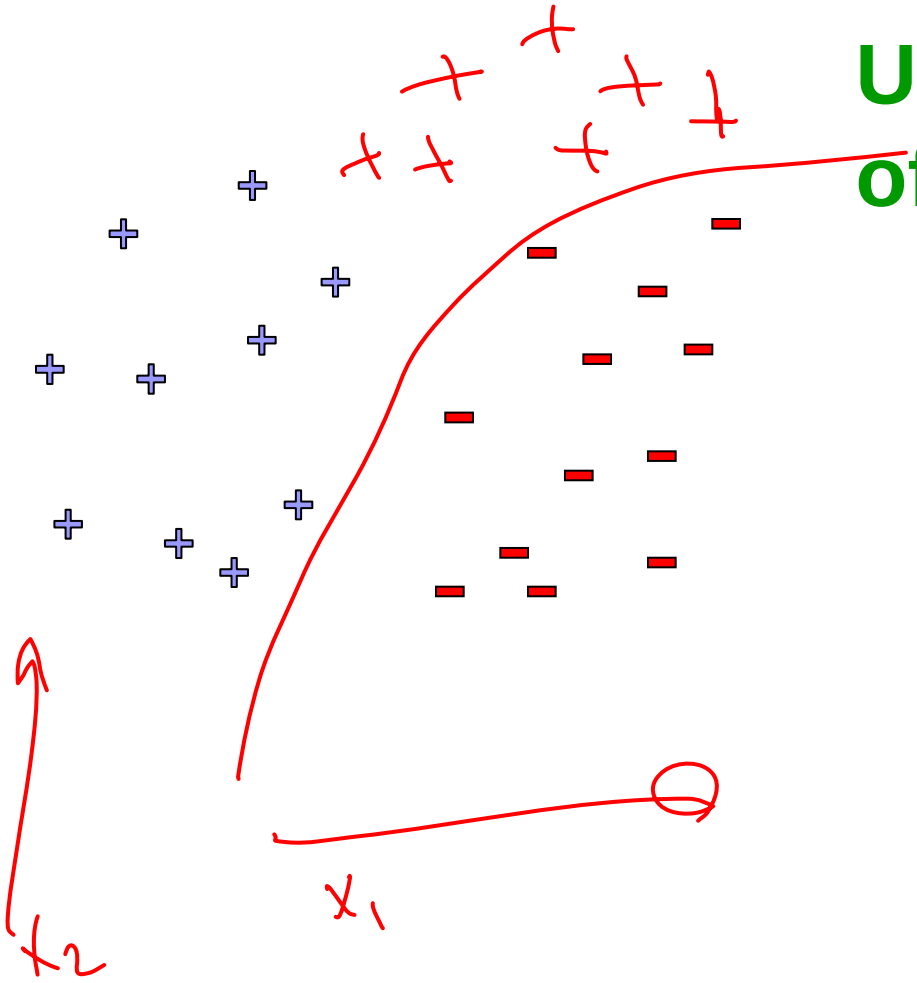
- Solve efficiently by quadratic programming (QP) → quadratic objective, linear constraints
  - Well-studied solution algorithms

- Hyperplane defined by support vectors

Announcements

- Third homework out later today
- This one is shorter!!!! :)
- Due on Monday March 5th
- No late days allowed
  - so we can give solutions before midterm

What if the data is not linearly separable?



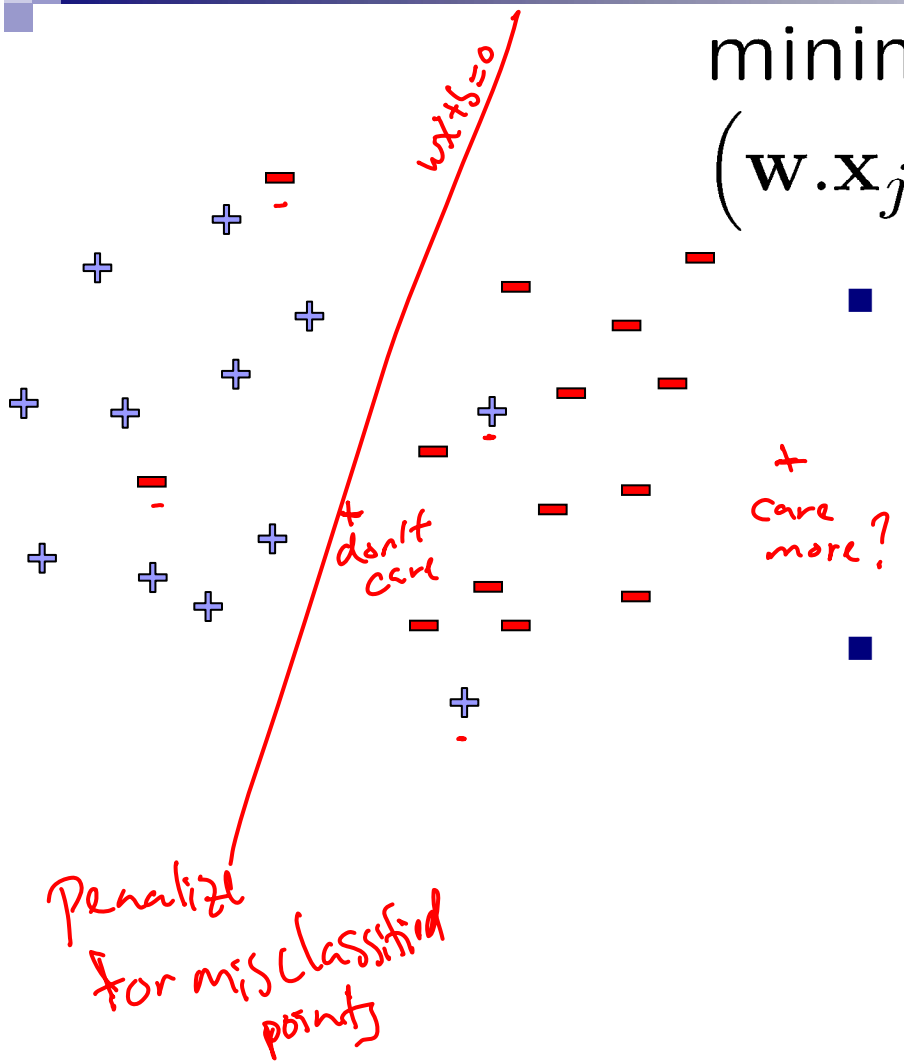
Use features of features of features of features….

features:

$\{x_1, x_2, x_1^2, x_2^2, x_1x_2, \ldots\}$

e.g., polynomials

What if the data is still not linearly separable?



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

magic constant

- Minimize **w.w** and number of training mistakes
  - Tradeoff two criteria?
  - C is tradeoff parameter
- Tradeoff #(mistakes) and **w.w**
  - 0/1 loss
  - Slack penalty C
  - Not QP anymore — optimization hard
  - Also doesn't distinguish near misses and really bad mistakes

$\delta(wx_j+b)y_j$

Slack variables – Hinge loss

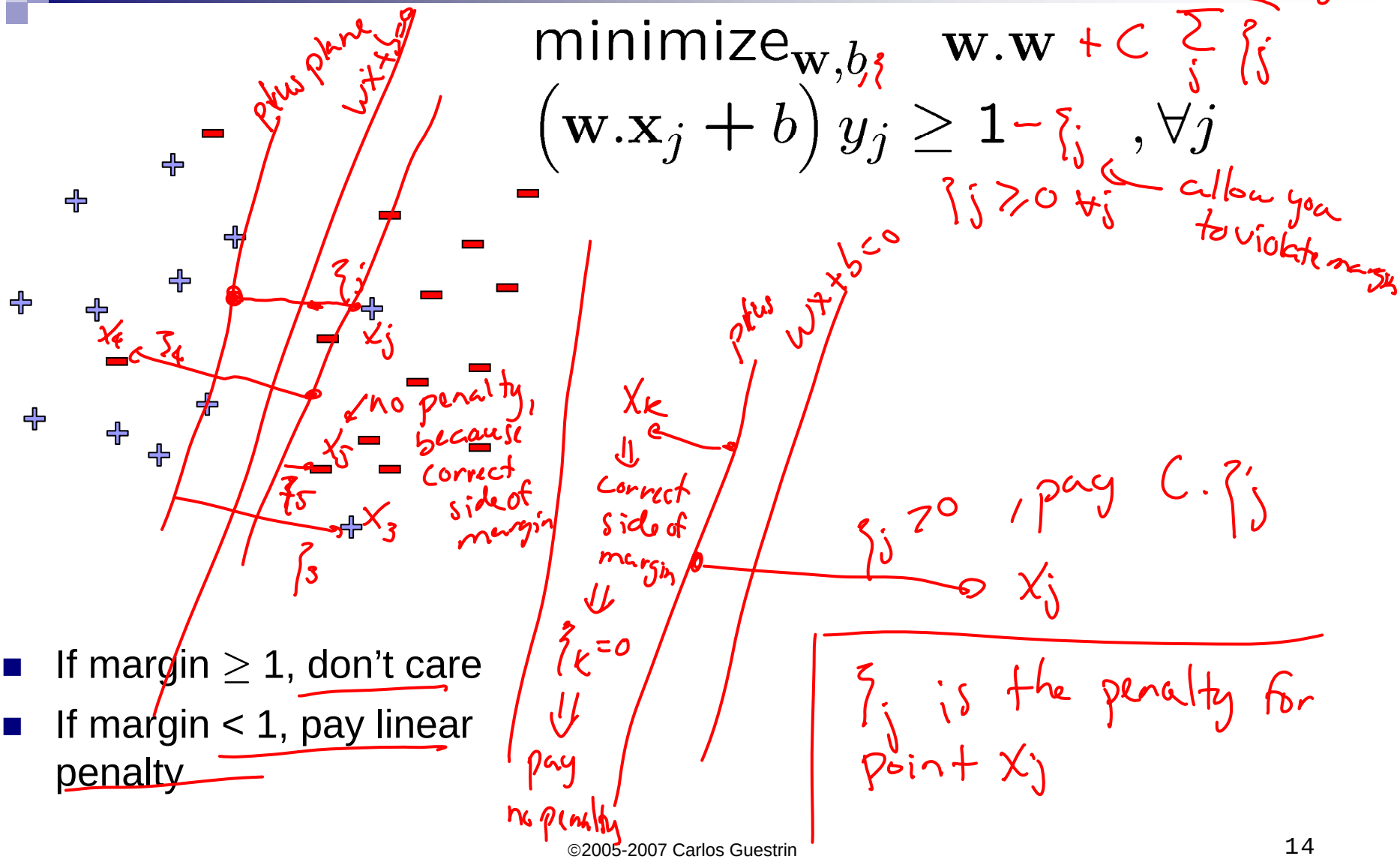
$$\text{minimize}_{\mathbf{w},b,\xi} \quad \mathbf{w}.\mathbf{w} + C\sum_j \xi_j$$

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

$\xi_j \geq 0 \ \forall j$

charge for violating margin

allow you to violate margin



- If margin ≥ 1, don't care
- If margin < 1, pay linear penalty

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

SVM:

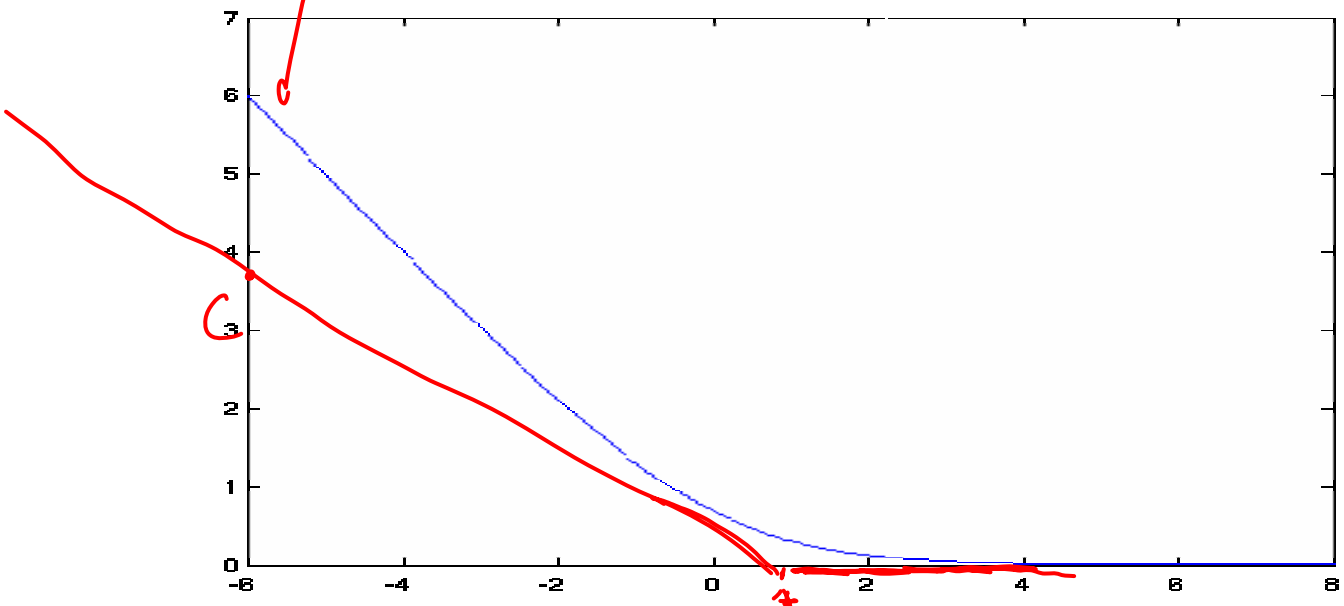
$$\text{minimize}_{\mathbf{w},b} \quad \mathbf{w}.\mathbf{w} + C\sum_j \xi_j$$
$$\left(\mathbf{w}.\mathbf{x}_j + b\right) y_j \geq 1 - \xi_j, \quad \forall j$$
$$\xi_j \geq 0, \quad \forall j$$

Logistic regression:

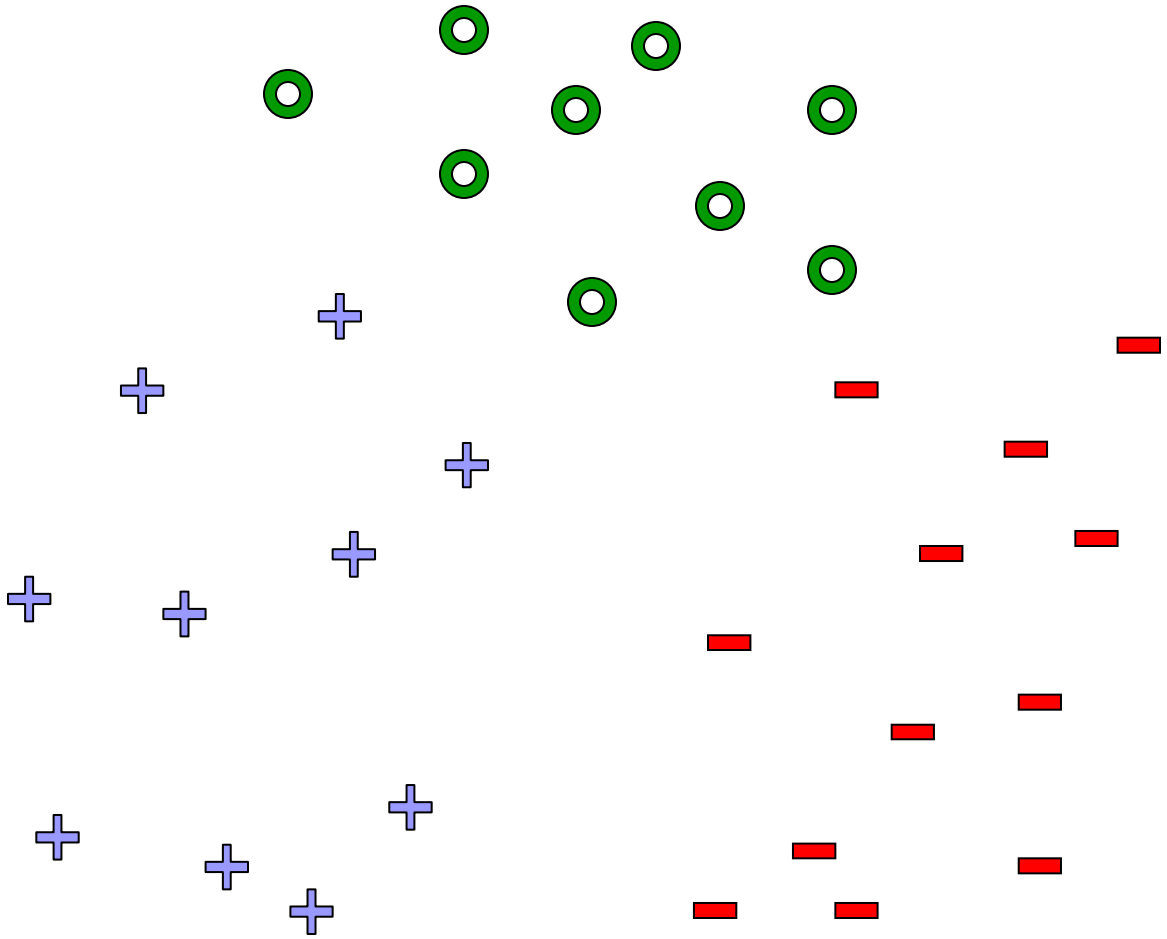
$$P(Y = 1 \mid x, \mathbf{w}, b) = \frac{1}{1 + e^{-(\mathbf{w}.\mathbf{x}+b)}}$$

Log loss:

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

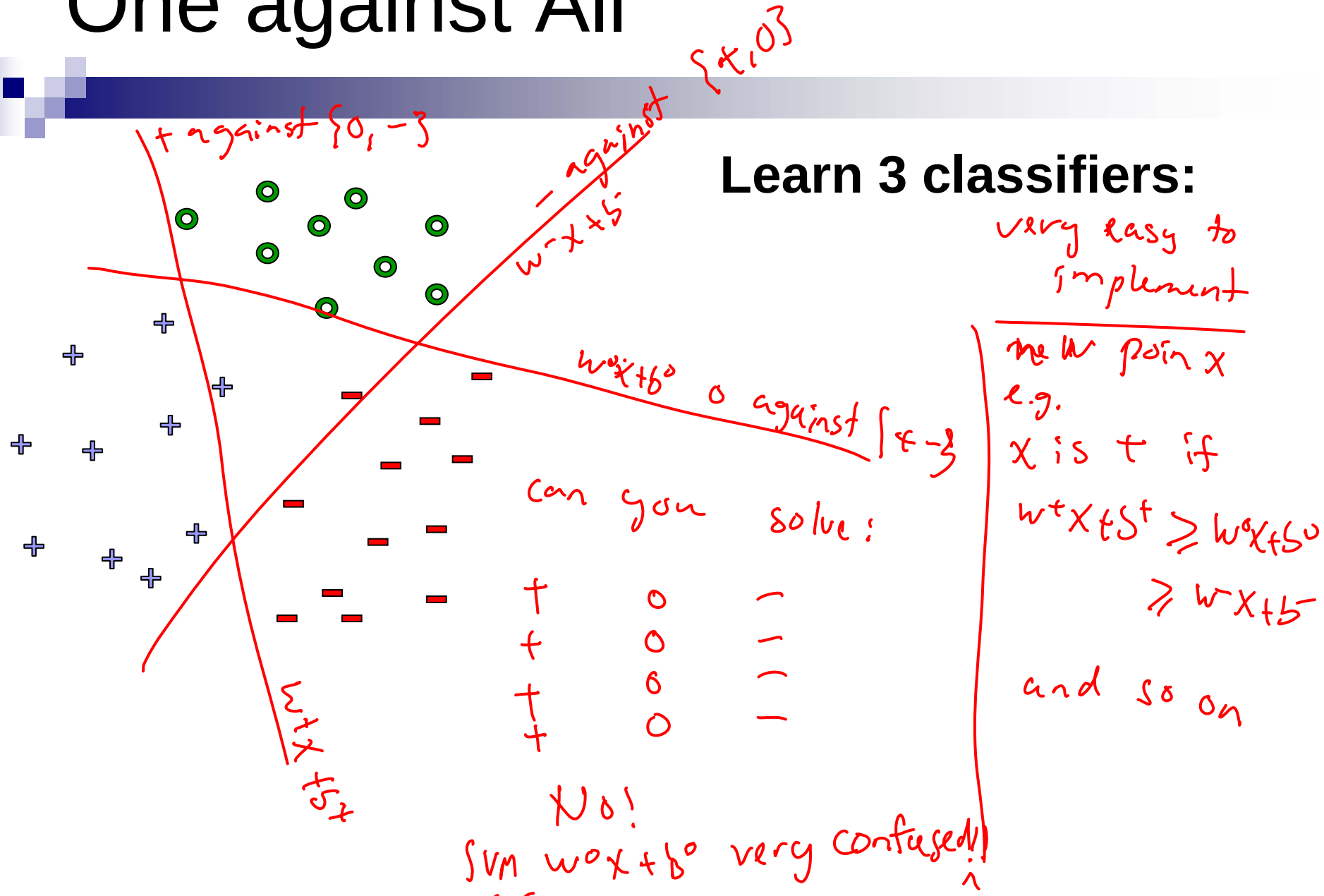


What about multiple classes?



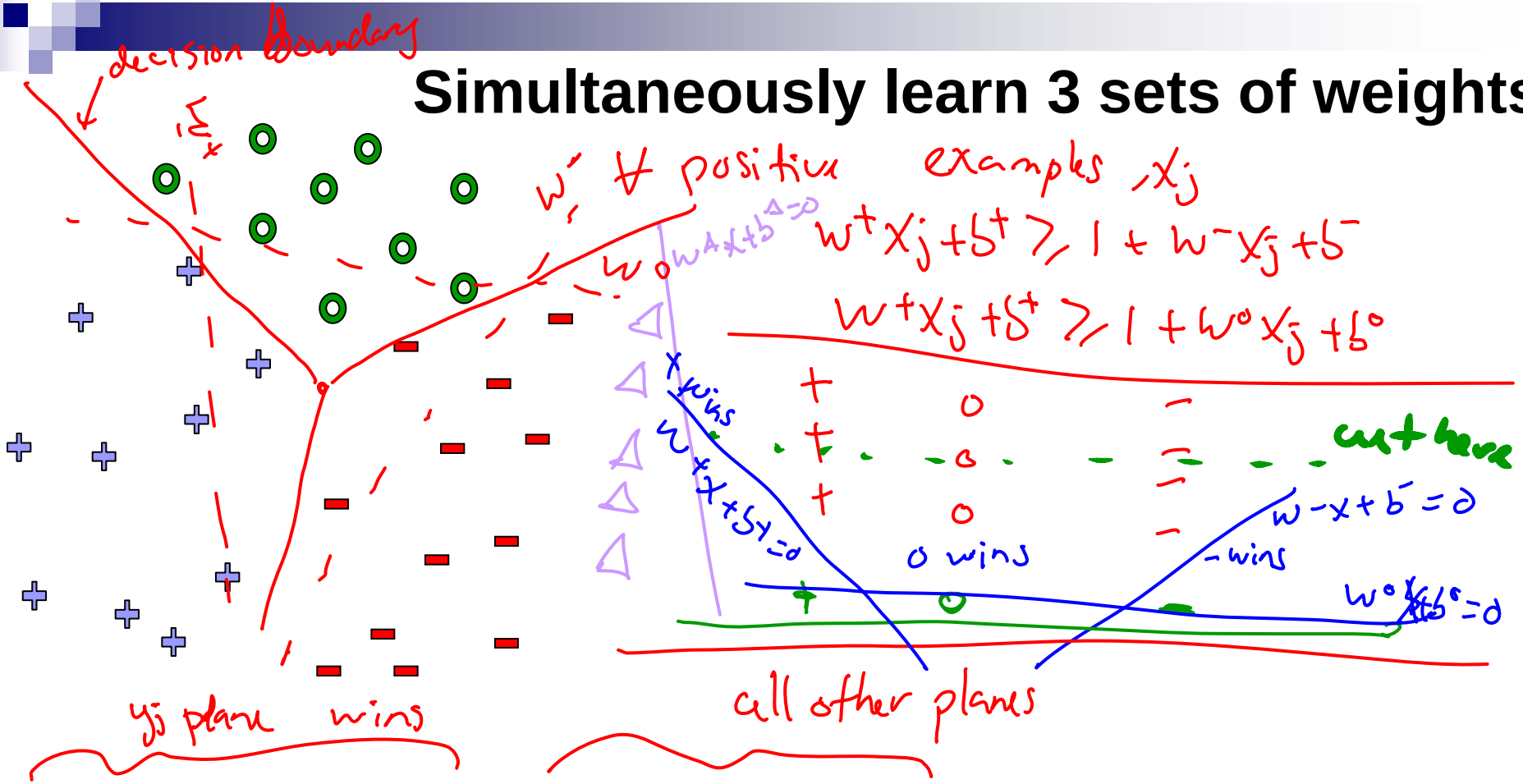
One against All

Learn 3 classifiers:



Learn 1 classifier: Multiclass SVM

Simultaneously learn 3 sets of weights



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

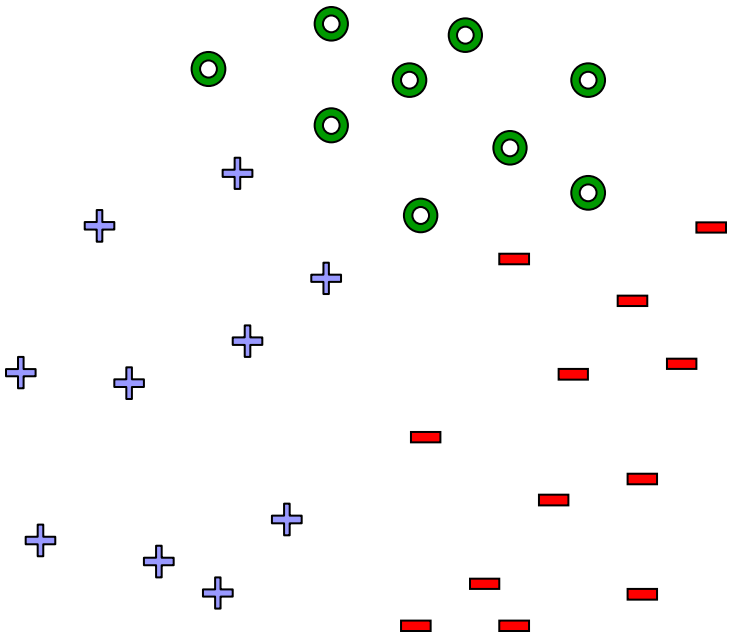
confidence of at least on

Learn 1 classifier: Multiclass SVM

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

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

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



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