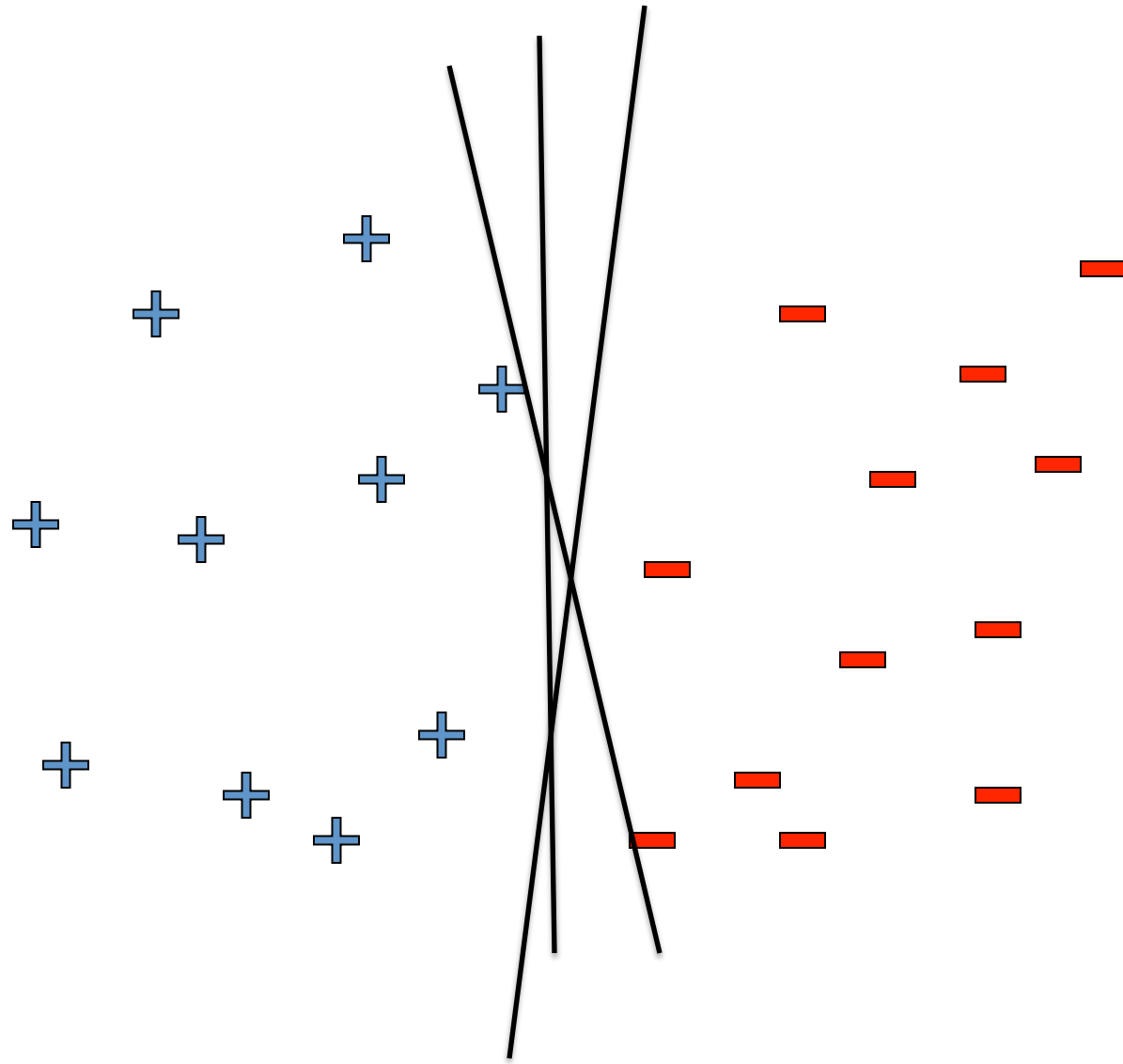


SVMs

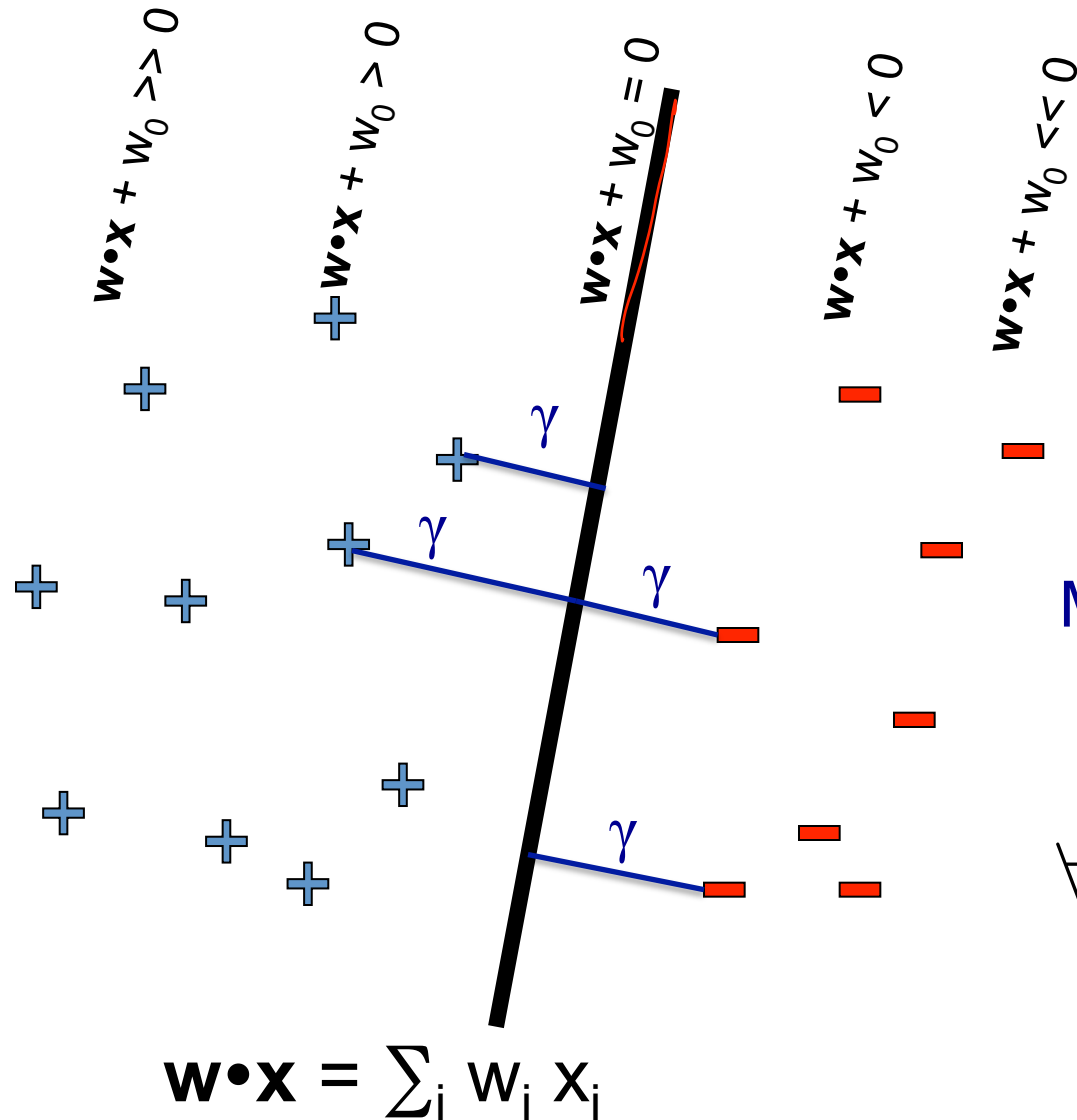
Instructor: Alan Ritter

Many Slides from Carlos Guestrin and Luke Zettlemoyer

Linear classifiers – Which line is better?



Pick the one with the largest margin!



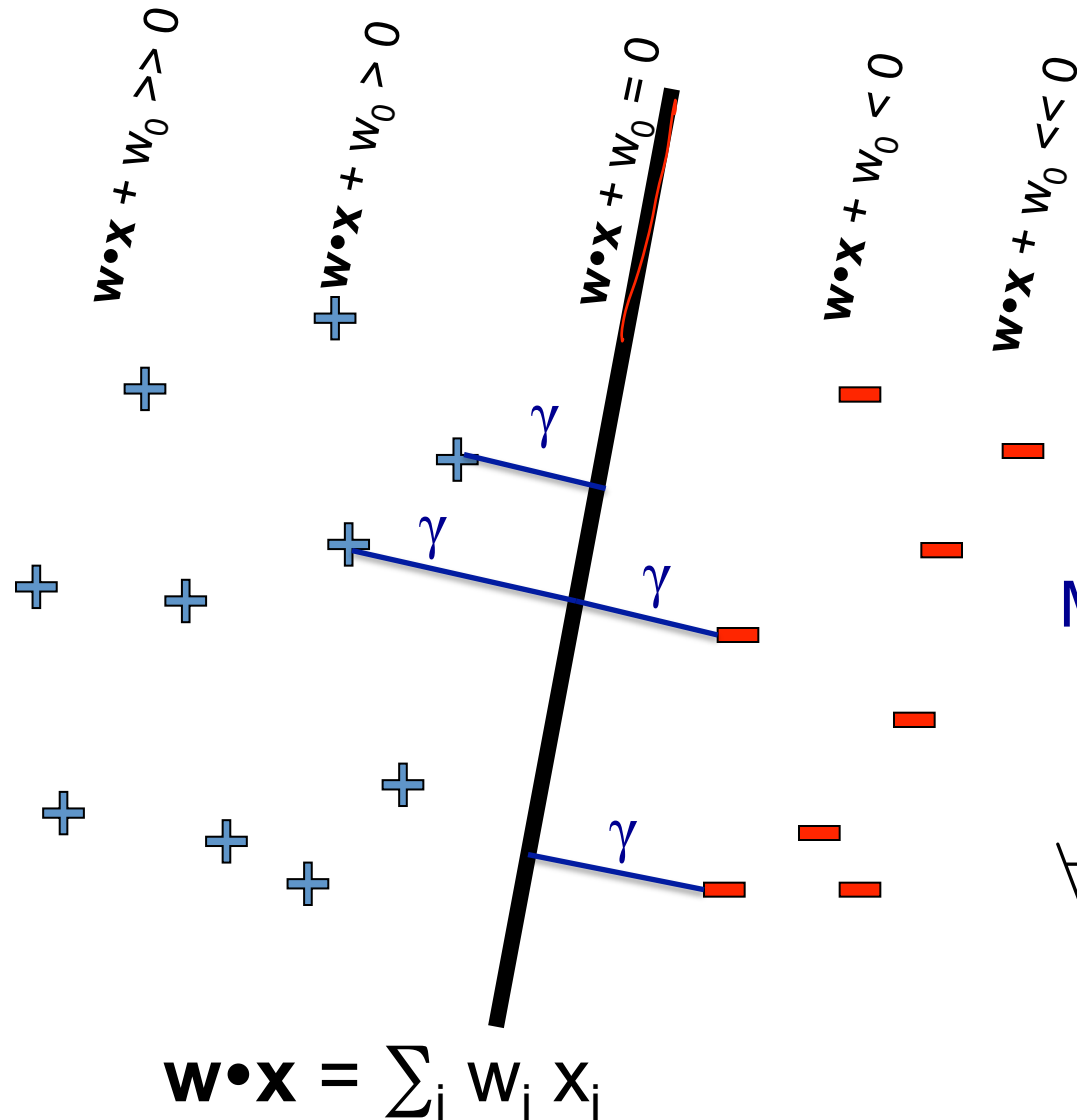
Margin for point j :

$$\gamma^j = y^j (w \cdot x^j + w_0)$$

Max Margin:

$$\max_{\gamma, w, w_0} \gamma$$
$$\forall j. y^j (w \cdot x^j + w_0) > \gamma$$

Pick the one with the largest margin!



Margin for point j:

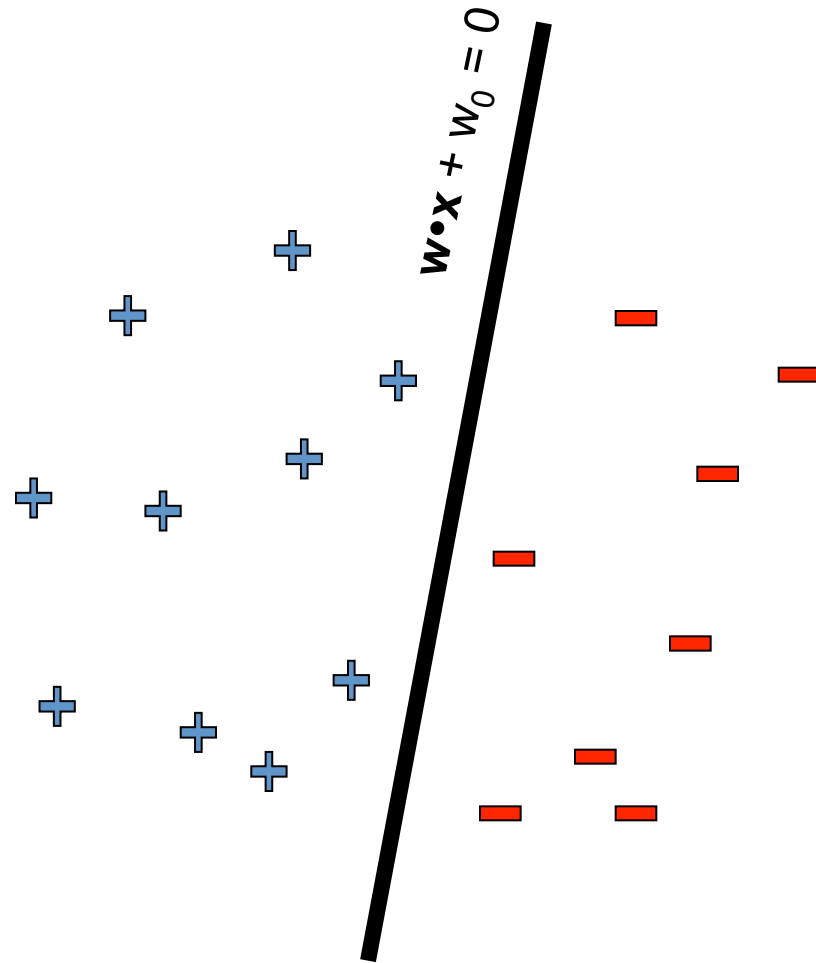
$$\gamma^j = y^j (w \cdot x^j + w_0)$$

Max Margin:

$$\begin{aligned} & \max_{\gamma, w, w_0} \gamma \\ & \forall j. y^j (w \cdot x^j + w_0) > \gamma \end{aligned}$$

Hard-margin
SVM

How many possible solutions?



$$\max_{\gamma, w, w_0} \gamma$$

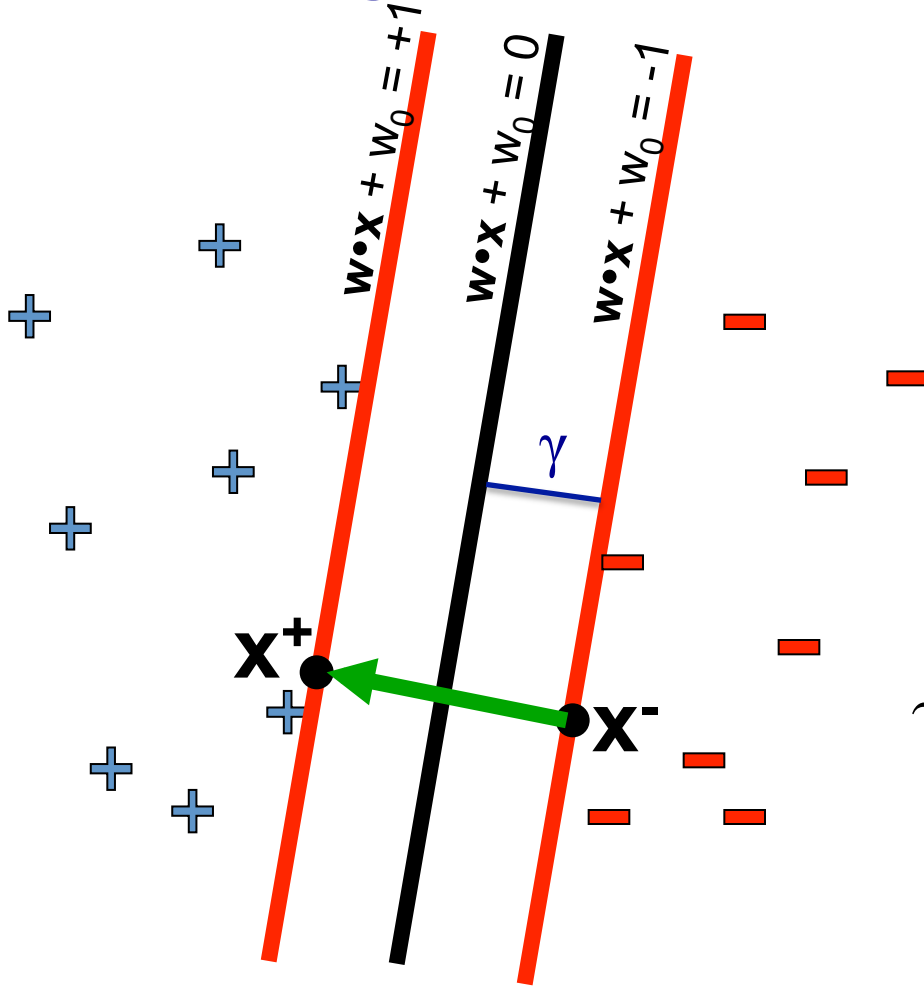
$$\forall j. y^j (w \cdot x^j + w_0) > \gamma$$

Any other ways of writing the same dividing line?

- $\mathbf{w} \cdot \mathbf{x} + b = 0$
- $2\mathbf{w} \cdot \mathbf{x} + 2b = 0$
- $1000\mathbf{w} \cdot \mathbf{x} + 1000b = 0$
-
- Any constant scaling has the same intersection with $z=0$ plane, so same dividing line!

Do we really want to \max_{γ, w, w_0} ?

Idea: *constrained*
margin



$$x^j = \bar{x}^j + \lambda \frac{w}{\|w\|_2}$$

$$\|w\|_2 = \sqrt{\sum_i w_i^2}$$

Assume: x^+ on positive line ($y=1$ intercept), x^- on negative ($y=-1$)

$$x^+ = x^- + 2\gamma \frac{w}{\|w\|}$$

$$w \cdot x^+ + w_0 = 1$$

$$w \cdot \left(x^- + 2\gamma \frac{w}{\|w\|_2} \right) + w_0 = 1$$

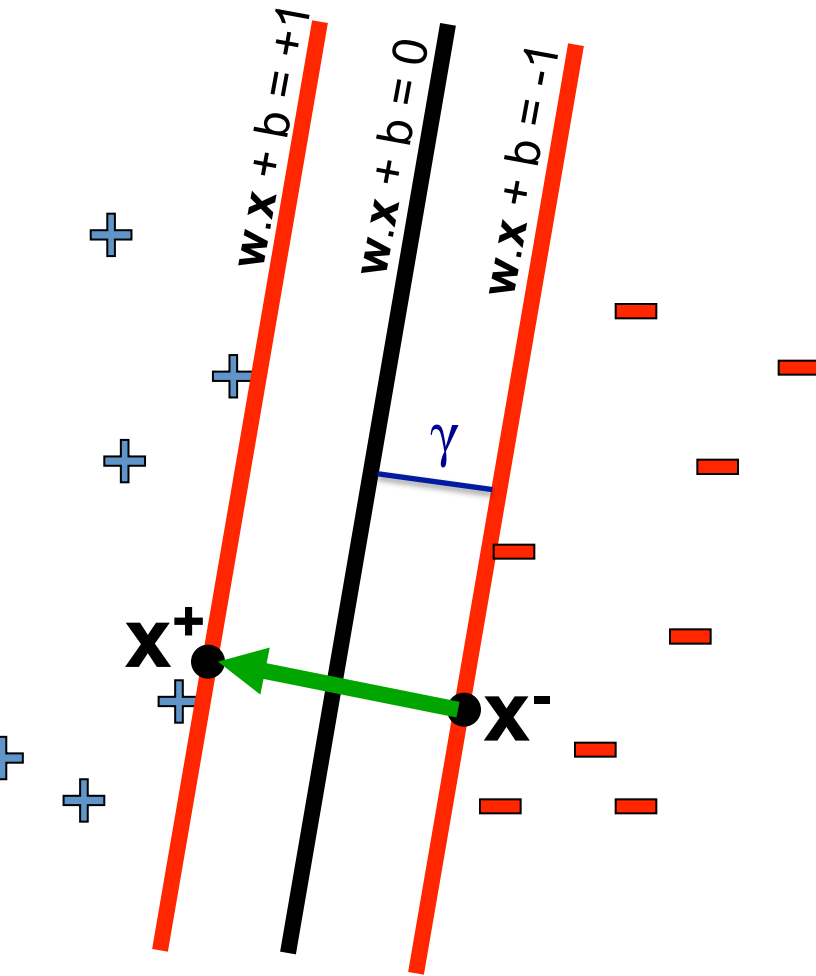
$$w \cdot x^- + w_0 + 2\gamma \frac{w \cdot w}{\|w\|_2} = 1$$

$$\gamma \frac{w \cdot w}{\|w\|_2} = 1 \quad w \cdot w = \sum_i w_i^2 = \|w\|_2^2$$

$$\gamma = \frac{\|w\|_2}{w \cdot w} = \frac{1}{\|w\|_2}$$

Final result: can maximize constrained margin by minimizing $\|w\|_2$!!!

Max margin using canonical hyperplanes



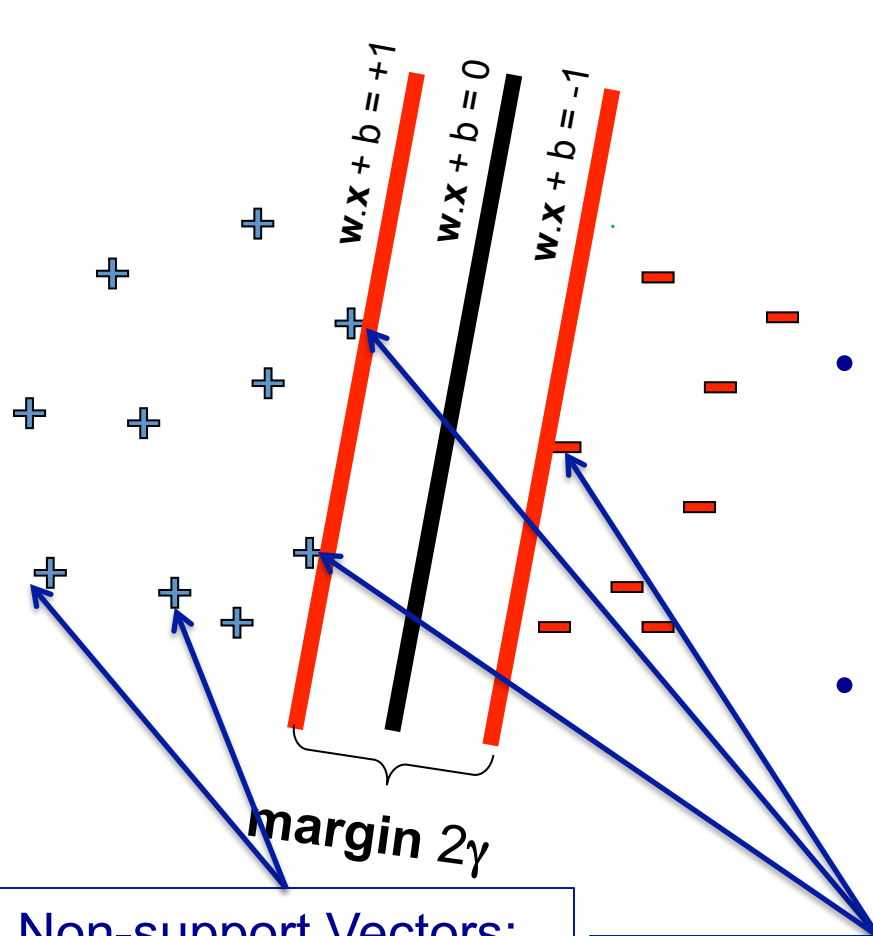
$$\max_{\gamma, w, w_0} \gamma$$
$$\forall j. y^j (w \cdot x^j + w_0) > \gamma$$

$$\gamma = \frac{1}{\|w\|_2}$$

$$\min_{w, w_0} \frac{1}{2} \|w\|_2^2$$
$$\forall j. y^j (w \cdot x^j + w_0) \geq 1$$

The assumption of canonical hyperplanes (at +1 and -1) changes the objective and the constraints!

Support vector machines (SVMs)



$$\min_{w, w_0} \frac{1}{2} \|w\|_2^2$$
$$\forall j. y^j (w \cdot x^j + w_0) \geq 1$$

- Solve efficiently by quadratic programming (QP)
 - Well-studied solution algorithms
 - Not simple gradient ascent, but close
- Decision boundary defined by support vectors

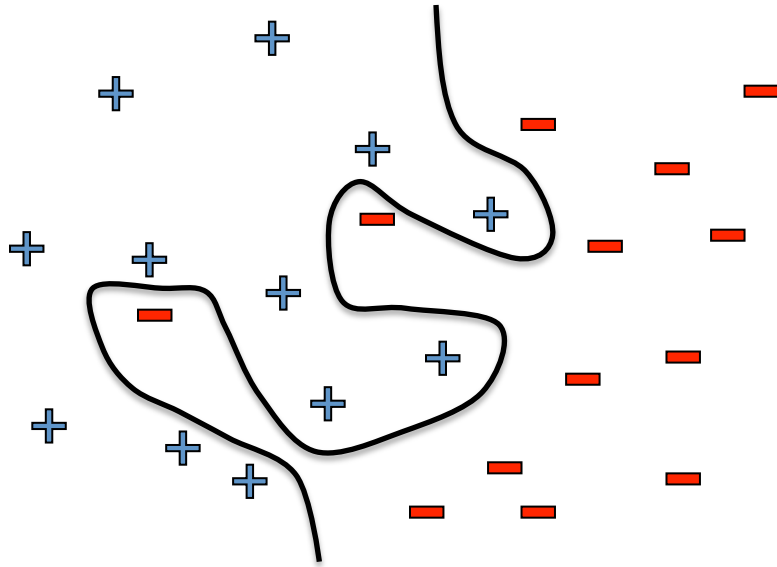
Non-support Vectors:

- everything else
- moving them will not change w

Support Vectors:

- data points on the canonical lines

What if the data is not linearly separable?



Add More Features!!!

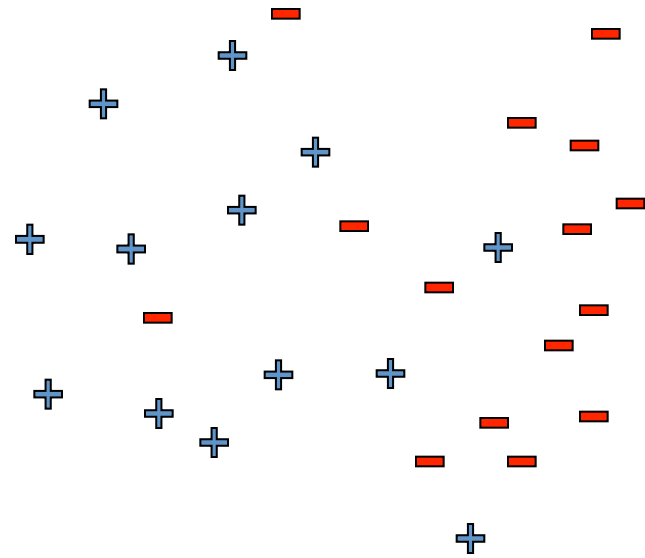
$$\phi(x) = \begin{pmatrix} x_1 \\ \dots \\ x_n \\ x_1 x_2 \\ x_1 x_3 \\ \dots \\ e^{x_1} \\ \dots \end{pmatrix}$$

Can use Kernels...
What about overfitting?

What if the data is still not linearly separable?

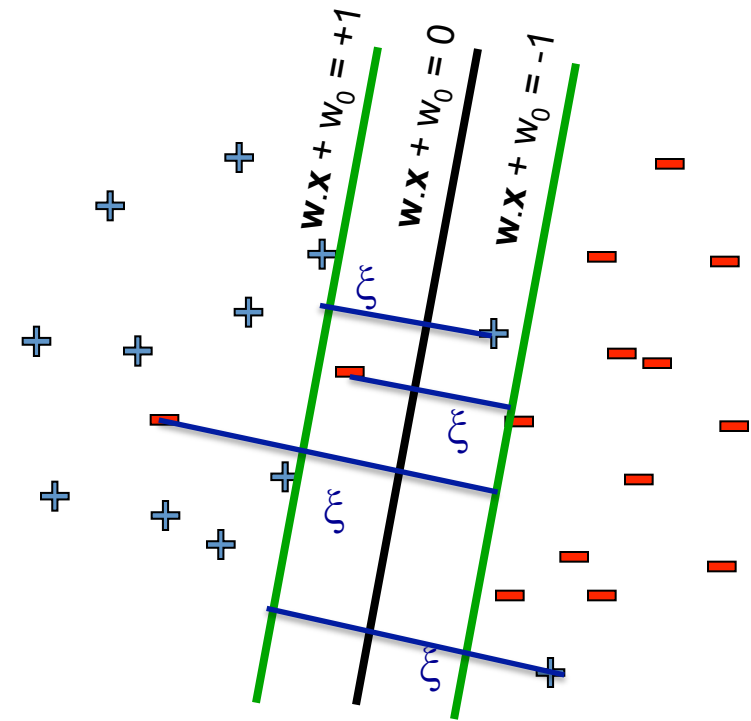
$$\min_{w, w_0} \frac{1}{2} \|w\|_2^2 + C \#(\text{mistakes})$$

$$\forall j. y^j (w \cdot x^j + w_0) \geq 1$$



- First Idea: Jointly minimize $\|w\|_2^2$ and number of training mistakes
 - How to tradeoff two criteria?
 - Pick C on development / cross validation
- Tradeoff $\#(\text{mistakes})$ and $\|w\|_2^2$
 - 0/1 loss
 - Not QP anymore
 - Also doesn't distinguish near misses and really bad mistakes
 - NP hard to find optimal solution!!!

Slack variables – Hinge loss



$$\min_{w, w_0} \frac{1}{2} \|w\|_2^2 + C \sum_j \xi_j$$
$$\forall j. y^j (w \cdot x^j + w_0) \geq 1 - \xi_j, \xi_j \geq 0$$

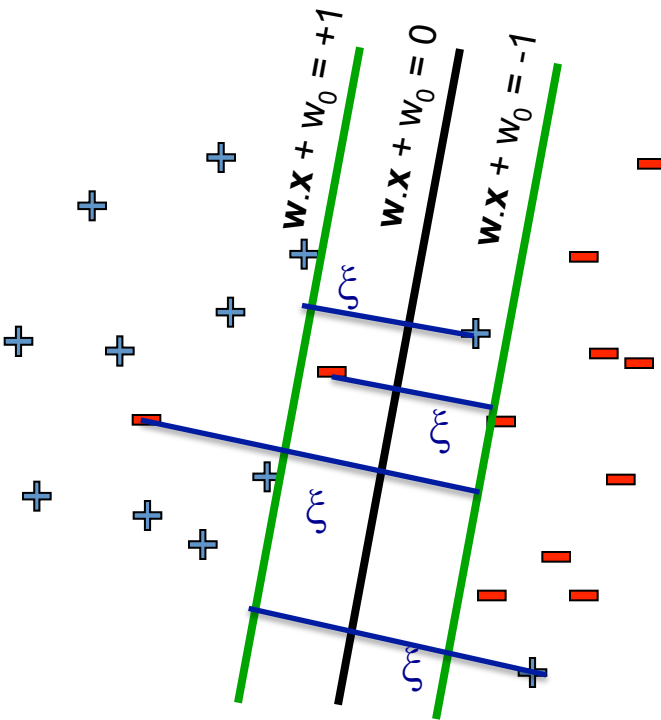
Slack Penalty $C > 0$:

- $C = \infty \rightarrow$ have to separate the data!
- $C = 0 \rightarrow$ ignore data entirely!
- Select on dev. set, etc.

For each data point:

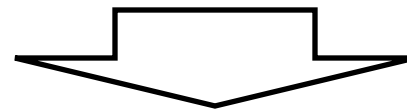
- If margin ≥ 1 , don't care
- If margin < 1 , pay linear penalty

Slack variables – Hinge loss



$$\min_{w, w_0} \frac{1}{2} \|w\|_2^2 + C \sum_j \xi_j$$

$$\forall j. y^j (w \cdot x^j + w_0) \geq 1 - \xi_j, \xi_j \geq 0$$



$$[x]_+ = \max(x, 0)$$

$$\min_{w, w_0} \frac{1}{2} \|w\|_2^2 + C \sum_{j=1}^N [1 - y^j (w \cdot x^j + w_0)]_+$$

Regularization

Hinge Loss

Solving SVMs:

- Differentiate and set equal to zero!
- No closed form solution, but quadratic program (top) is concave
- Hinge loss is not differentiable, gradient ascent a little trickier...

Logistic Regression as Minimizing Loss

Logistic regression assumes:

$$f(x) = w_0 + \sum_i w_i x_i$$

$$P(Y = 1 | X = x) = \frac{\exp(f(x))}{1 + \exp(f(x))}$$

And tries to maximize data likelihood, for $Y = \{-1, +1\}$:

$$P(y^i | x^i) = \frac{1}{1 + \exp(-y^i f(x^i))}$$
$$\begin{aligned} \ln P(\mathcal{D}_Y | \mathcal{D}_X, \mathbf{w}) &= \sum_{j=1}^N \ln P(y^j | \mathbf{x}^j, \mathbf{w}) \\ &= - \sum_{i=1}^N \ln(1 + \exp(-y^i f(x^i))) \end{aligned}$$

Equivalent to minimizing *log loss*:

$$\sum_{i=1}^N \ln(1 + \exp(-y^i f(x^i)))$$

SVMs vs Regularized Logistic Regression

$$f(x) = w_0 + \sum_i w_i x_i$$

SVM Objective:

$$\arg \min_{\mathbf{w}, w_0} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{j=1}^N [1 - y^j f(x^j)]_+$$

$$[x]_+ = \max(x, 0)$$

Logistic regression objective:

$$\arg \min_{\mathbf{w}, w_0} \lambda \|\mathbf{w}\|_2^2 + \sum_{j=1}^N \ln(1 + \exp(-y^j f(x^j)))$$

Tradeoff: same l_2 regularization term, but different error term

Graphing Loss vs Margin

Logistic regression:

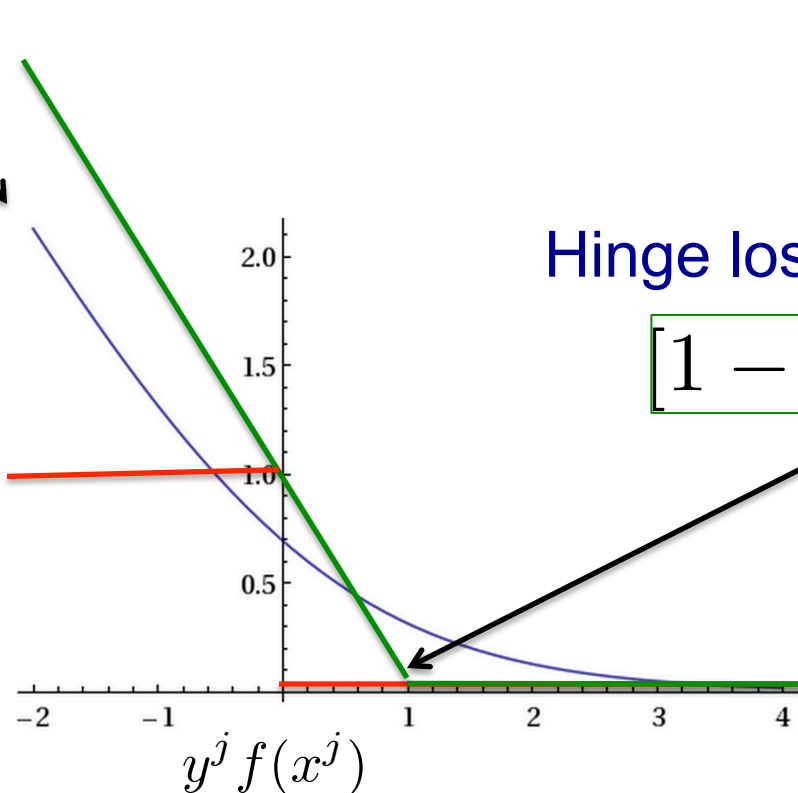
$$\ln(1 + \exp(-y^j f(x^j)))$$

Hinge loss:

$$[1 - y^j f(x^j)]_+$$

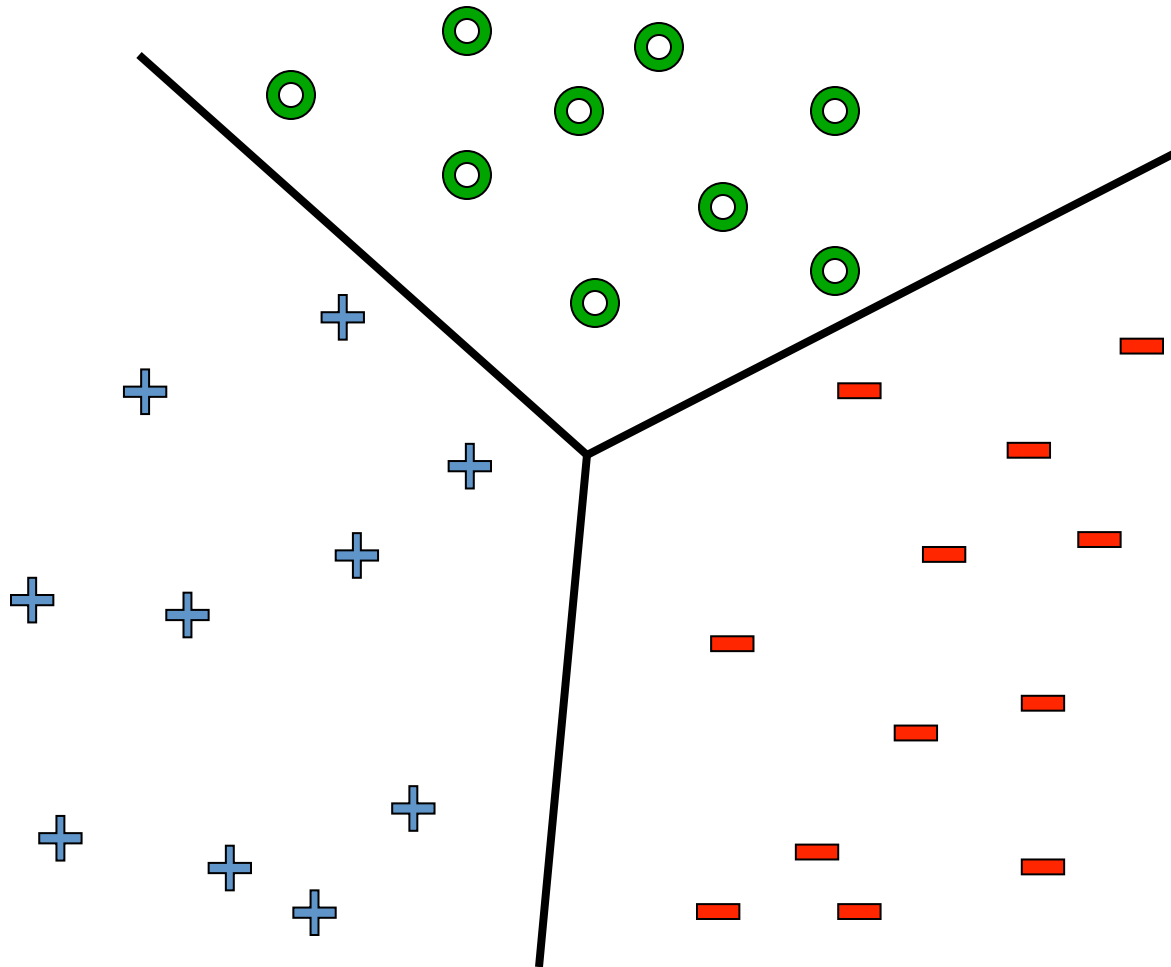
0-1 Loss:

$$\delta(f(x^j) \neq y^j)$$

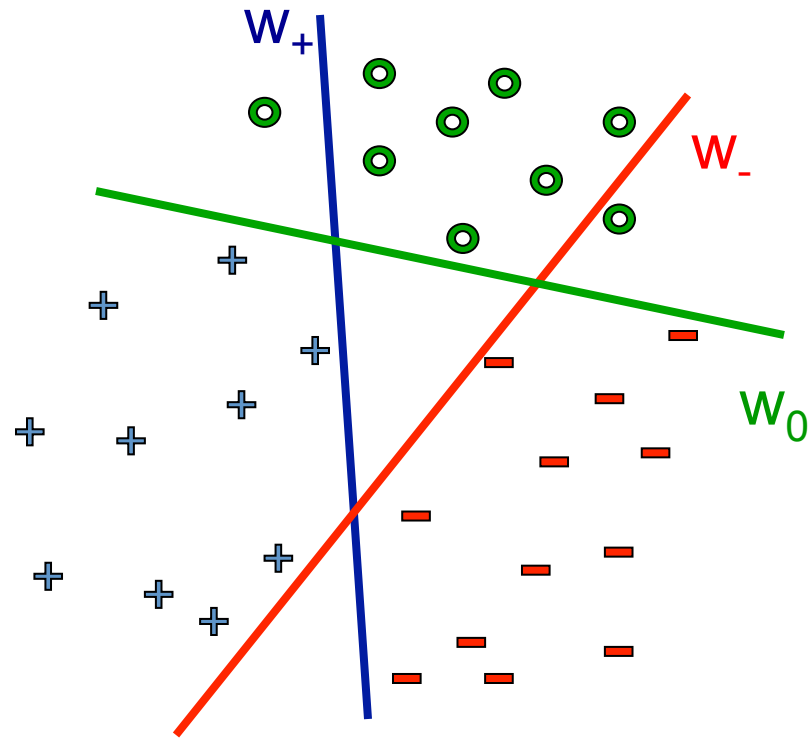


We want to smoothly approximate 0/1 loss!

What about multiple classes?



One against All



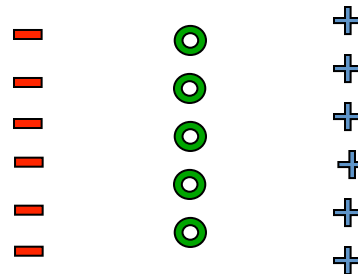
Learn 3 classifiers:

- + vs {0,-}, weights w_+
- - vs {0,+}, weights w_-
- 0 vs {+,-}, weights w_0

Output for x :

$$y = \operatorname{argmax}_i w_i \bullet x$$

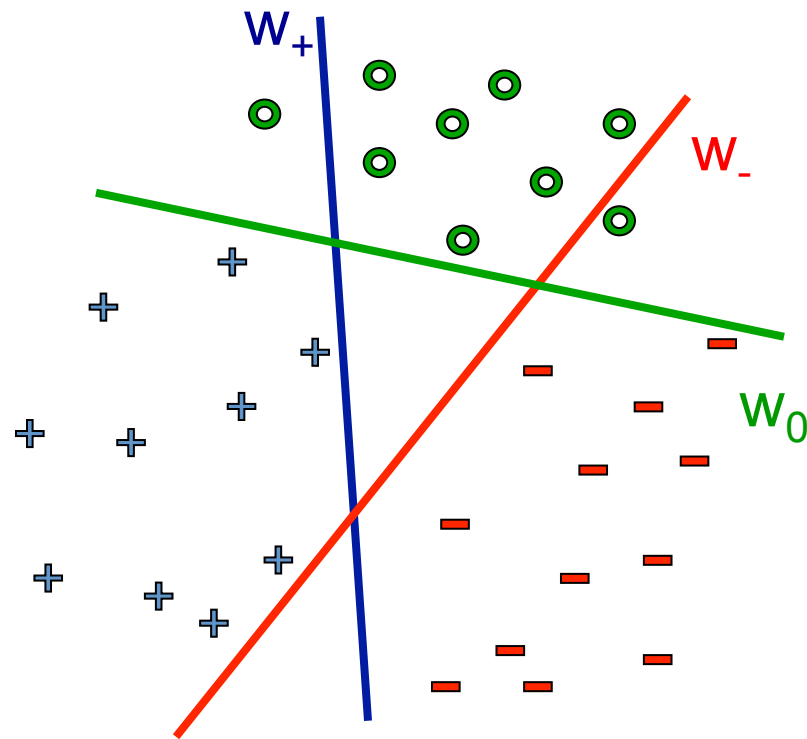
Any problems?
Could we learn this \rightarrow
dataset?



Learn 1 classifier: Multiclass SVM

Simultaneously learn 3 sets of weights:

- How do we guarantee the correct labels?
- Need new constraints!



For each class:

$$w^{y^j} \cdot x^j + w_0^{y^j} \geq w^{y'} \cdot x^j + w_0^{y'} + 1, \quad \forall y' \neq y^j, \quad \forall j$$

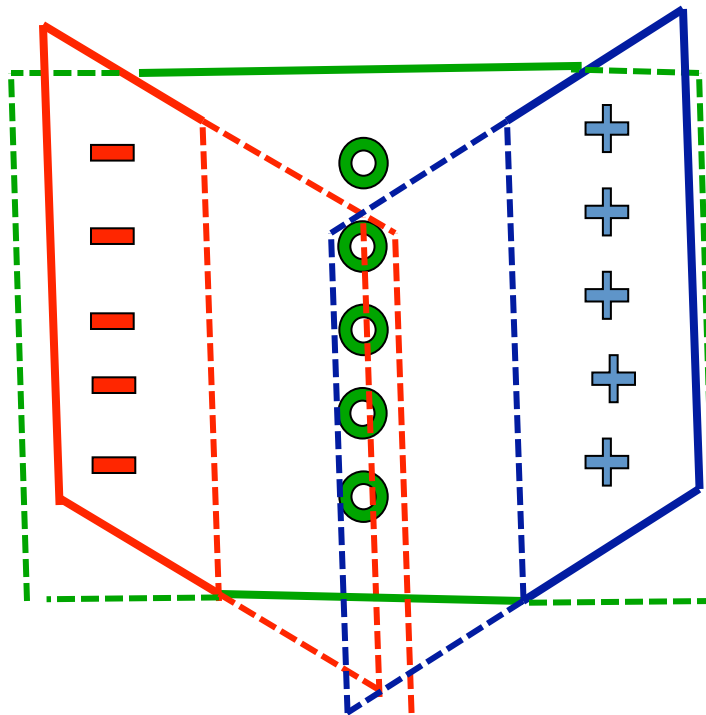
Learn 1 classifier: Multiclass SVM

Also, can introduce slack variables, as before:

$$\min_{w, w_0} \sum_y \|w^y\|_2^2 + C \sum_j \xi^j$$

$$w^{y^j} \cdot x^j + w_0^{y^j} \geq w^{y'} \cdot x^j + w_0^{y'} + 1 - \xi^j, \quad \forall y' \neq y^j, \quad \xi^j > 0 \quad \forall j$$

Now, can we learn it?



Regularization

$$f(x) = w_0 + \sum_i w_i x_i$$

SVM Objective:

$$\arg \min_{\mathbf{w}, w_0} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{j=1}^N [1 - y^j f(x^j)]_+$$

$$[x]_+ = \max(x, 0)$$

Logistic regression objective:

$$\arg \min_{\mathbf{w}, w_0} \lambda \|\mathbf{w}\|_2^2 + \sum_{j=1}^N \ln(1 + \exp(-y^j f(x^j)))$$

L1 Regularization

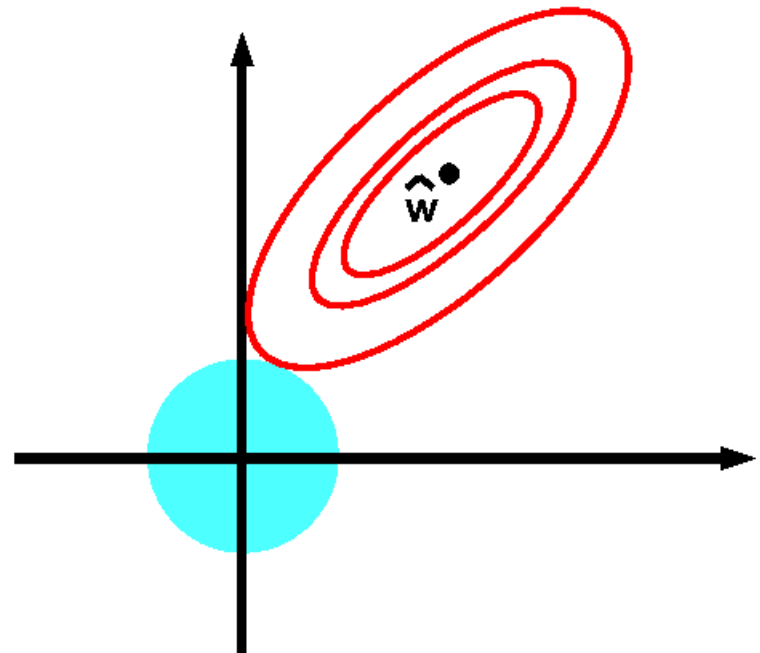
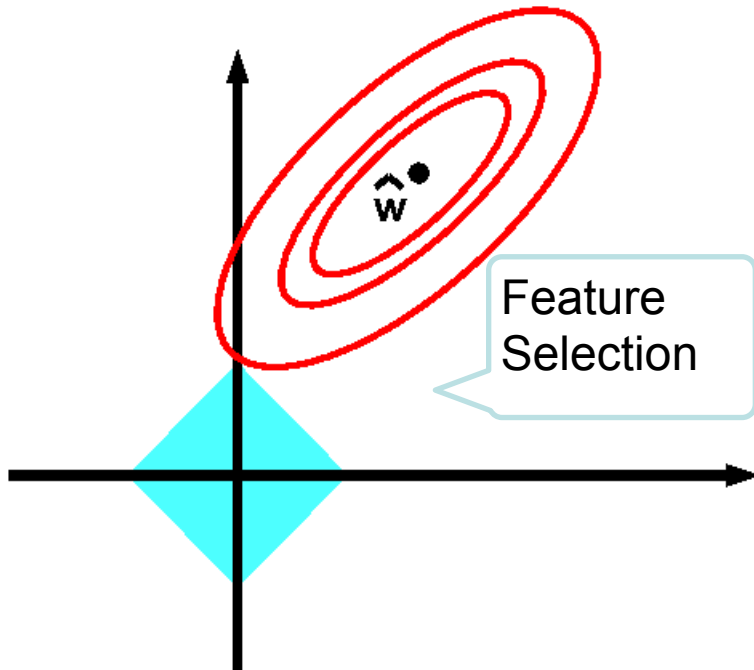
$$\arg \min_{\mathbf{w}, w_o} \lambda \|\mathbf{w}\|_1 + \sum_{j=1}^N \ln(1 + \exp(-y^j f(x^j)))$$

$$\|w\|_1 = \sum_i |w_i|$$

L1 vs. L2 Regularization

$$||w||_1 = \sum_i |w_i|$$

$$||w||_2^2 = \sum_i w_i^2$$



What you need to know

- Maximizing margin
- Derivation of SVM formulation
- Slack variables and hinge loss
- Tackling multiple class
 - One against All
 - Multiclass SVMs