Support Vector Machine

Foundations of Data Analysis

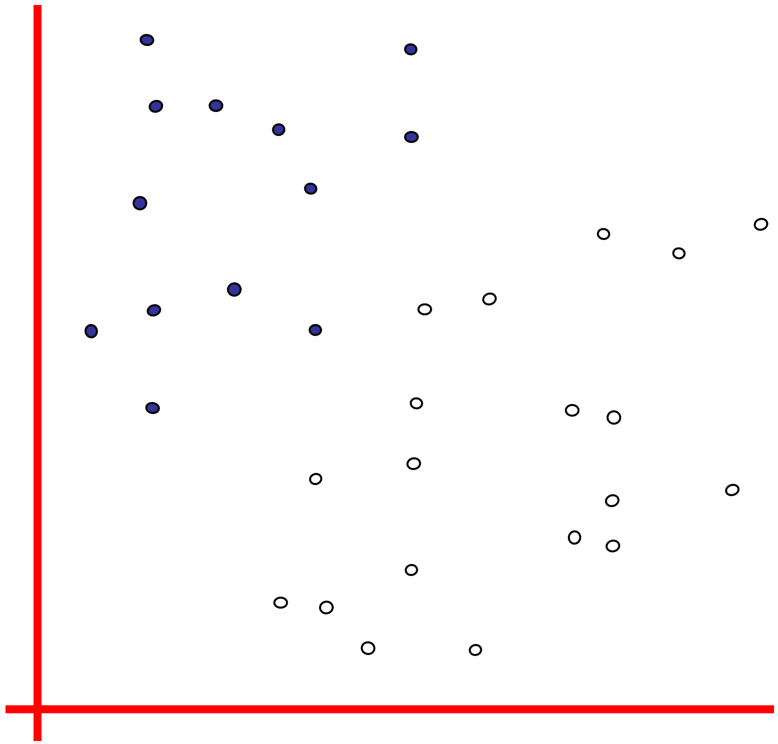
04/14/2020

Slide credits to Andrew Moore: http://www.cs.cmu.edu/~awm/tutorials

Linear Classifiers

$$f(x, w) = \text{sign}(w \cdot x)$$

- denotes +1
- denotes -1

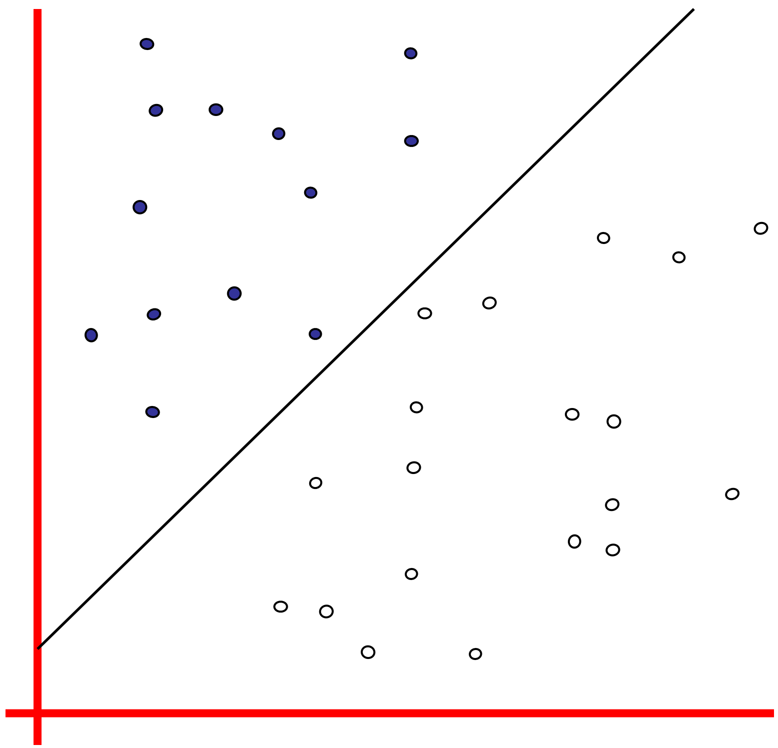


How would you classify this data?

Linear Classifiers

$$f(x, w) = \mathrm{sign}(w \cdot x)$$

- denotes +1
- denotes -1

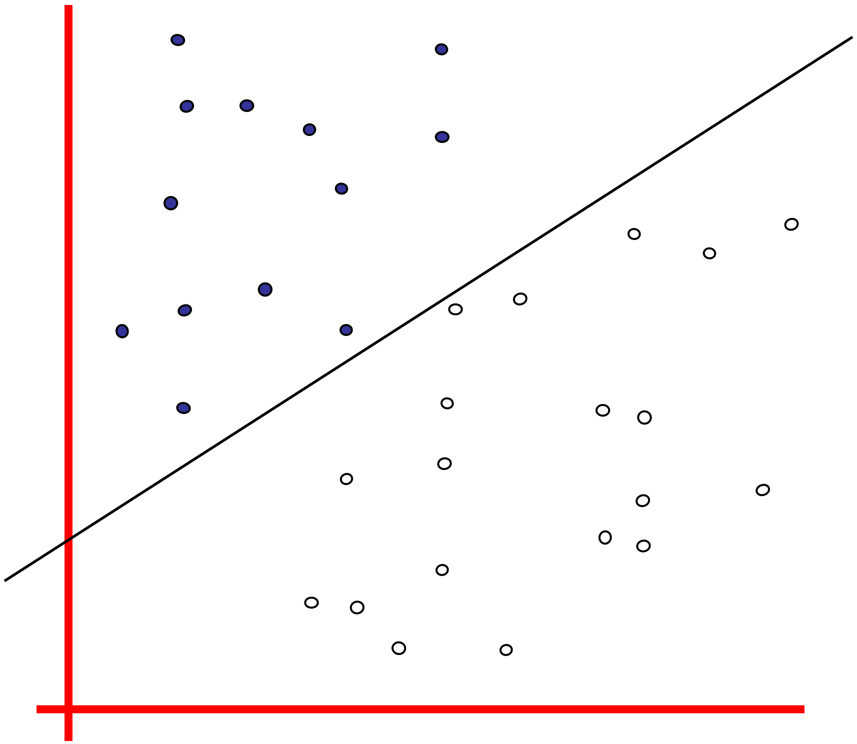


How would you classify this data?

Linear Classifiers

$$f(x, w) = \mathrm{sign}(w \cdot x)$$

- denotes +1
- denotes -1

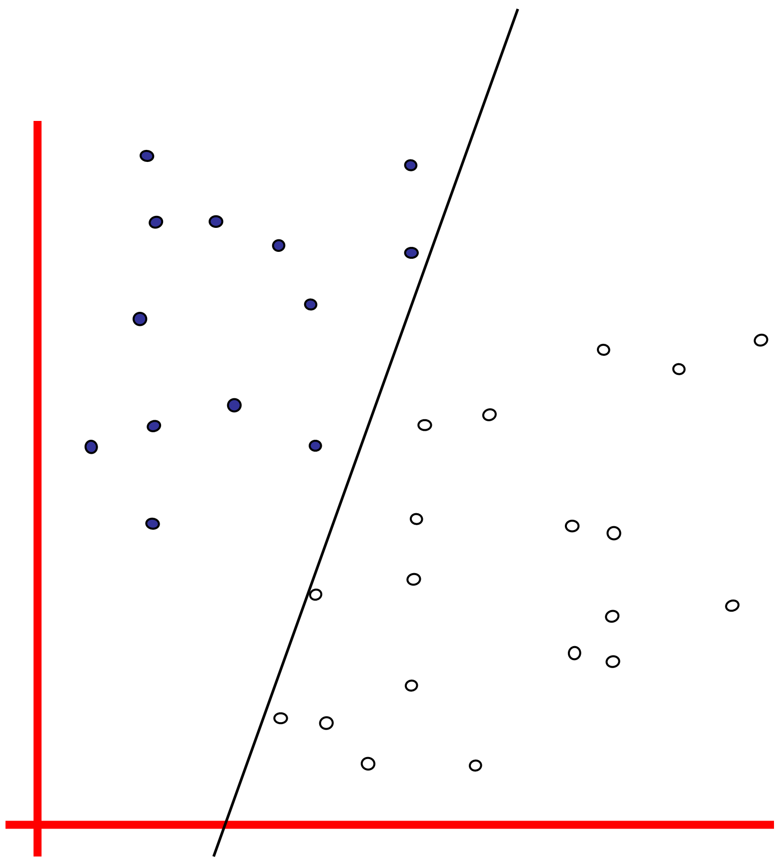


How would you classify this data?

Linear Classifiers

$$f(x, w) = \mathrm{sign}(w \cdot x)$$

- denotes +1
- denotes -1

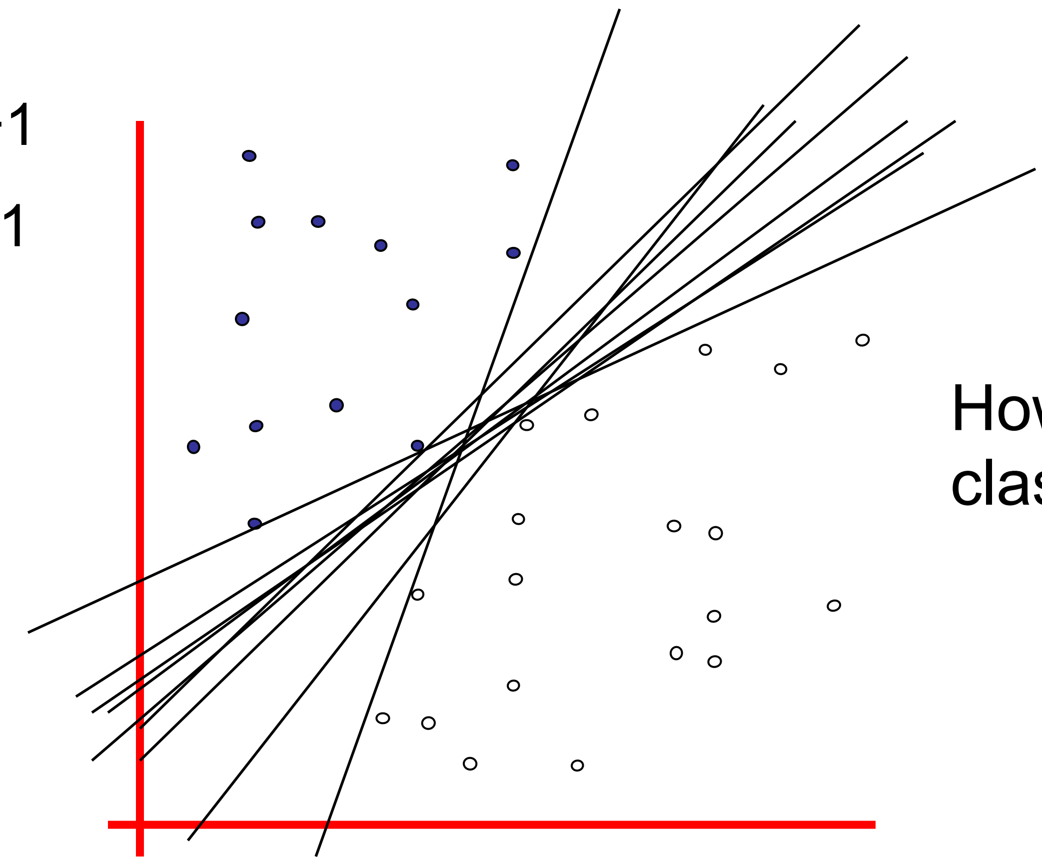


How would you classify this data?

Linear Classifiers

$$f(x, w) = \mathrm{sign}(w \cdot x)$$

- • denotes +1
- ◦ denotes -1

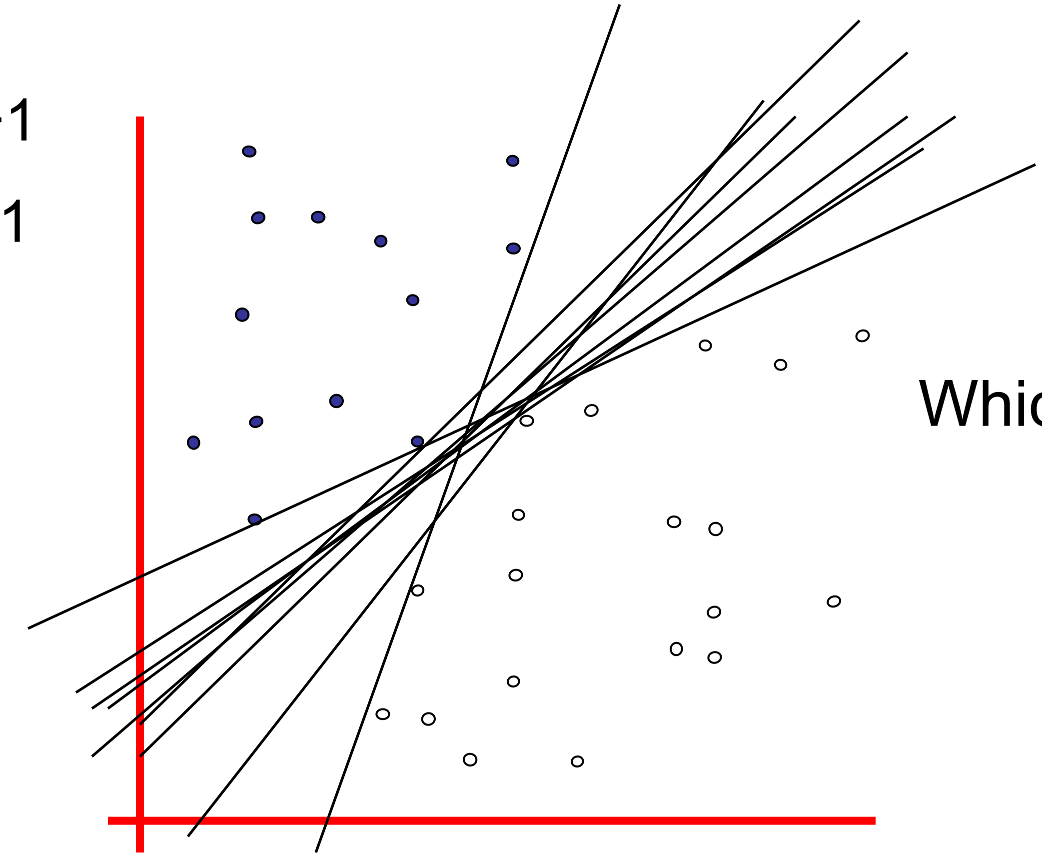


How would you classify this data?

Linear Classifiers

$$f(x, w) = \mathrm{sign}(w \cdot x)$$

- denotes +1
- denotes -1

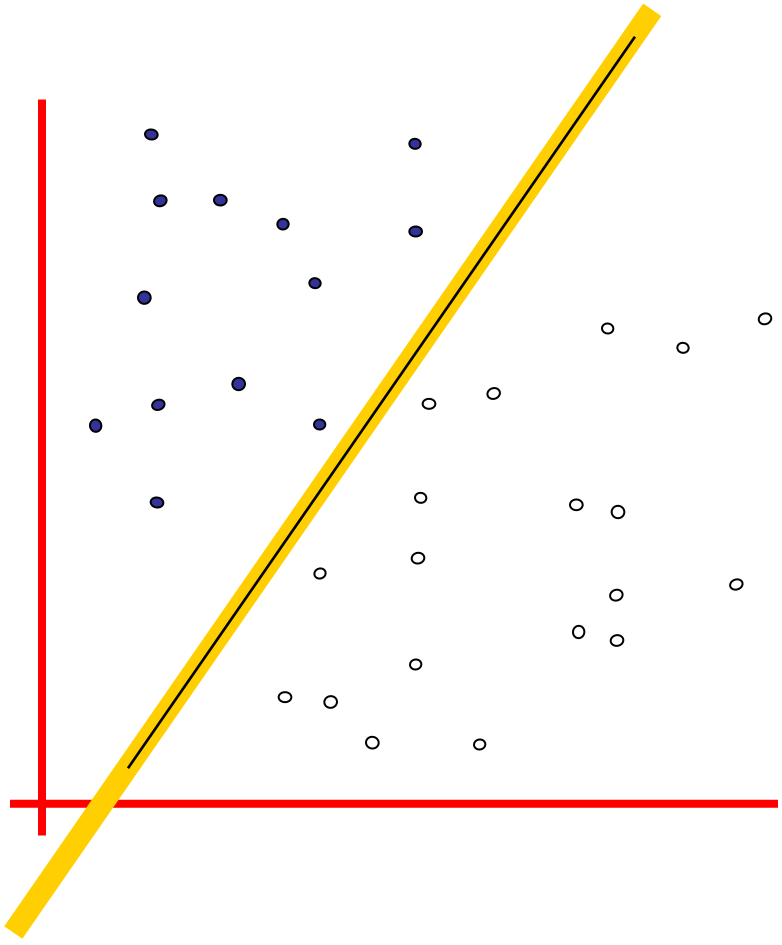


Which is the best?

Linear Classifiers

$$f(x, w) = \text{sign}(w \cdot x)$$

- denotes +1
- denotes -1

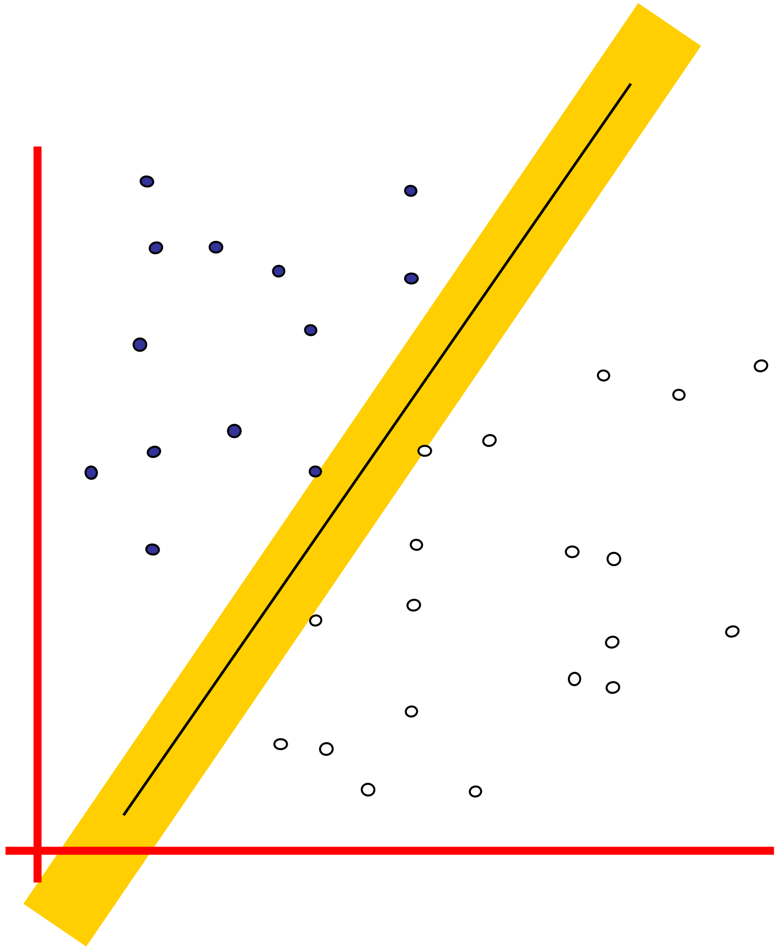


Define the margin of a linear classifier as the width that the boundary could be increased before hitting a datapoint.

Maximum Margin

$$f(x, w) = \text{sign}(w \cdot x)$$

- denotes +1
- denotes -1

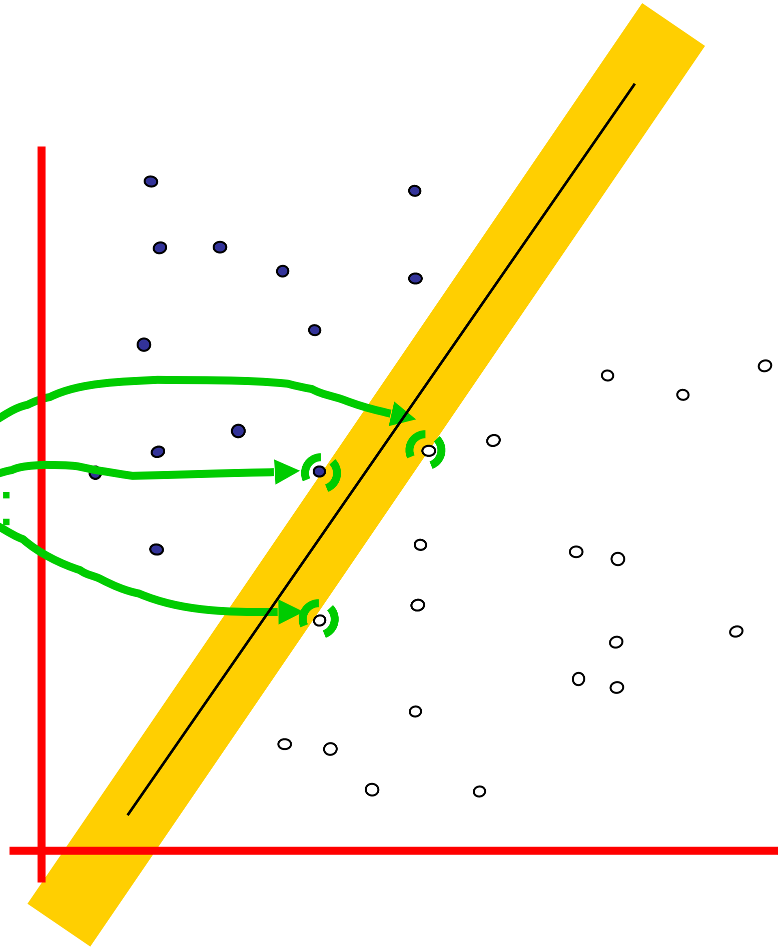


Maximum margin: the widest margin that maximally separates two data groups.

Support Vector Machines

$$f(x, w) = \mathrm{sign}(w \cdot x)$$

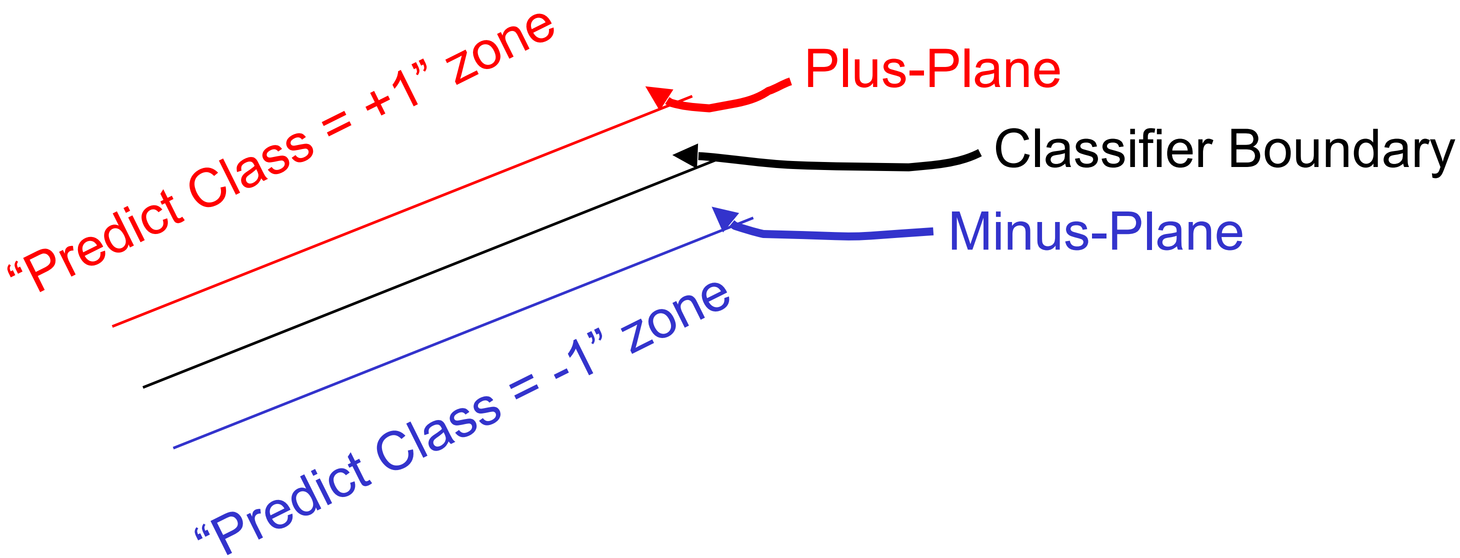
- denotes +1
- denotes -1



Support Vectors : data points that the margin pushes up against.

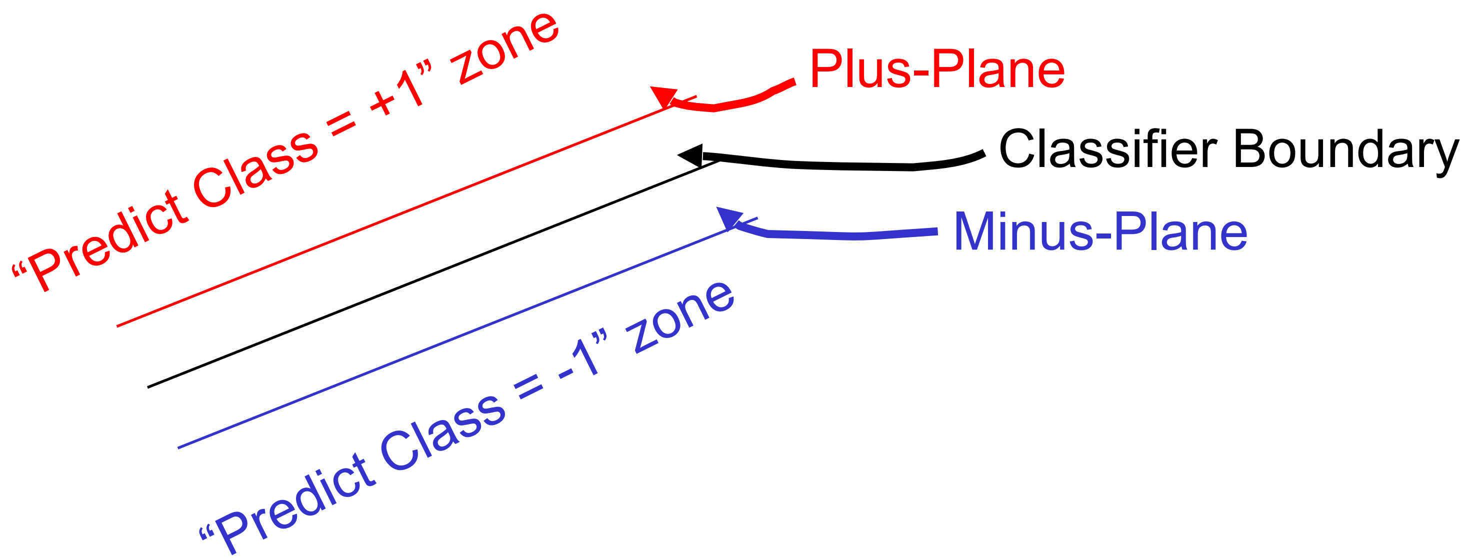
Maximum margin linear classifier is the simplest kind of SVM (LinearSVM)

Specifying a line and margin



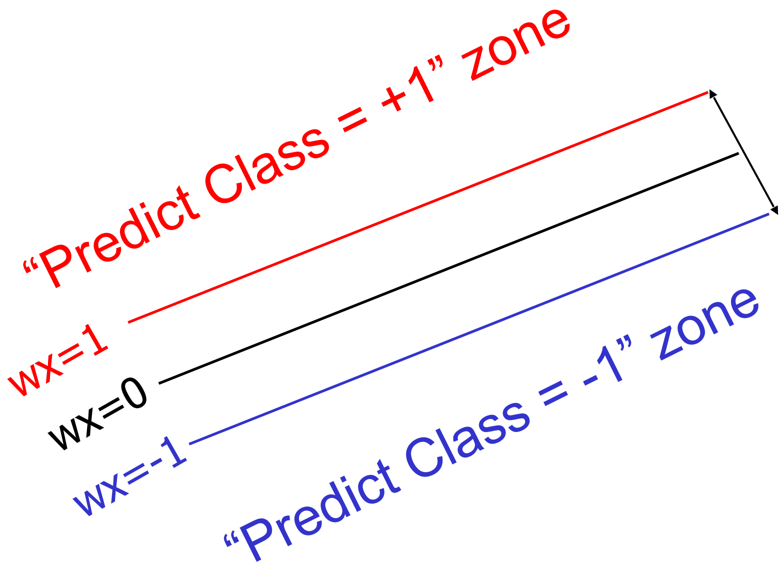
- How do we represent this mathematically?
- …in m input dimensions?

Specifying a line and margin



Class	+1	if	$w \cdot x >= 1$
	-1	if	$w \cdot x <= -1$
	embarrassing points	if	$-1 < w \cdot x < 1$

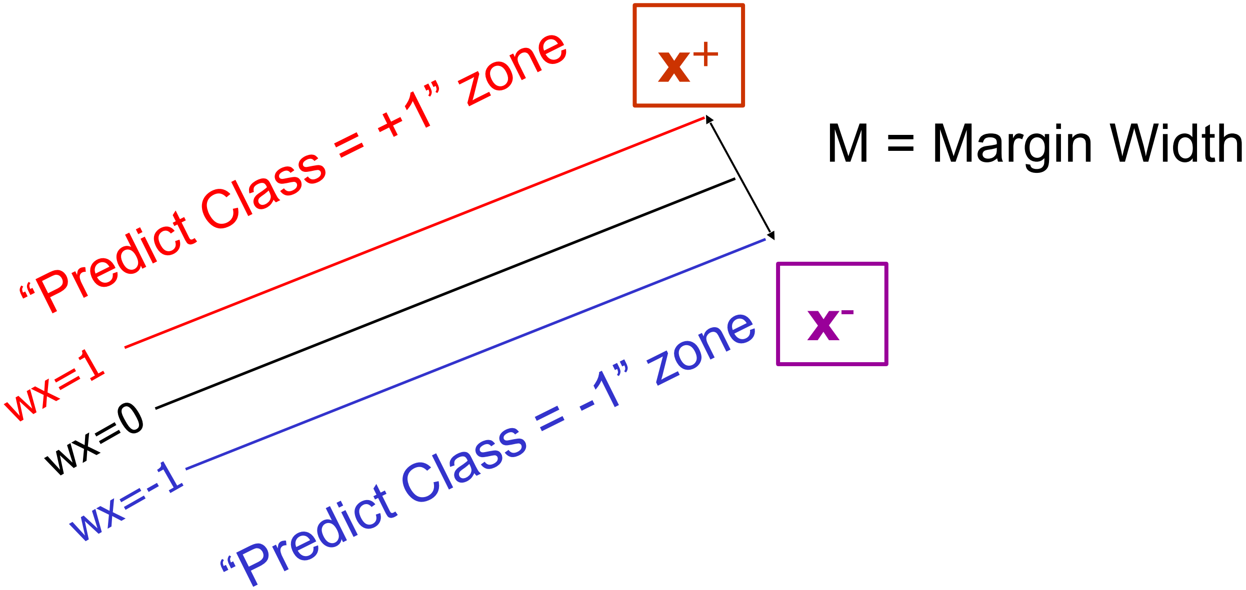
Computing the Margin Width



M = Margin Width

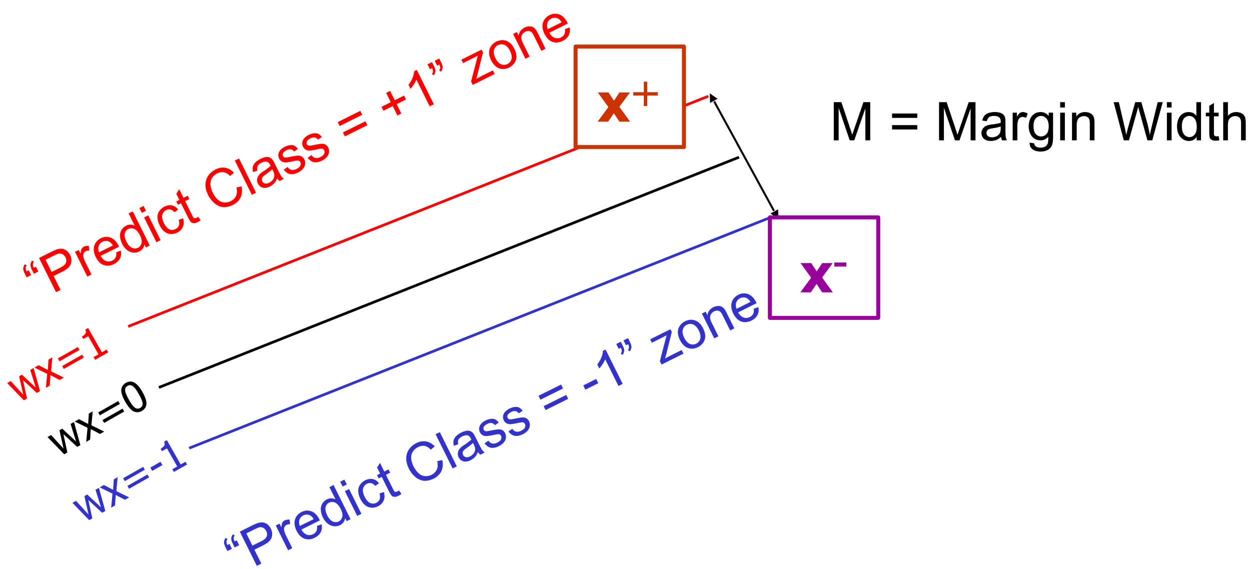
How to compute M?

Computing the Margin Width



- Let $\mathbf{x}^-$ be any point on the minus plane
- Let $\mathbf{x}^+$ be the closest plus-plane-point to $\mathbf{x}^-$
- Claim: $\mathbf{x}^+ = \mathbf{x}^- + \lambda\, \mathbf{w}$ for some value of λ. Why?

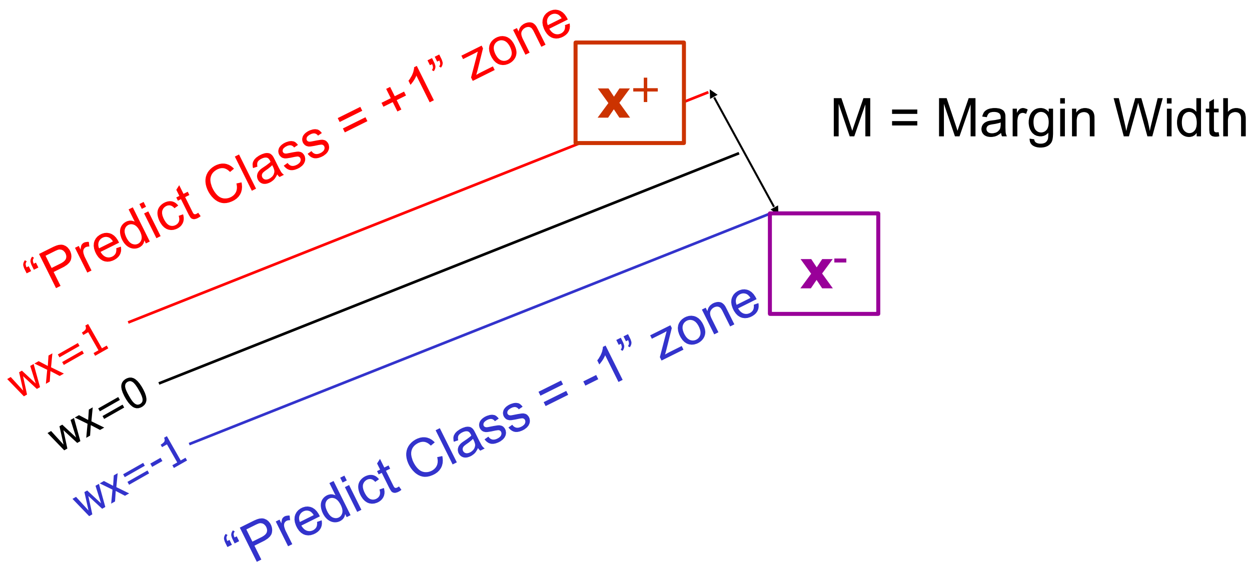
Computing the Margin Width



- Let $\mathbf{x}^-$ be any point on the minus plane
- Let $\mathbf{x}^+$ be the closest plus-plane-point to $\mathbf{x}^-$
- Claim: $\mathbf{x}^+ = \mathbf{x}^- + \lambda \, \mathbf{w}$ for some value of λ. Why?

The line from $\mathbf{x}^-$ to $\mathbf{x}^+$ is perpendicular to the planes. So to get from $\mathbf{x}^-$ to $\mathbf{x}^+$ travel some distance in direction w.

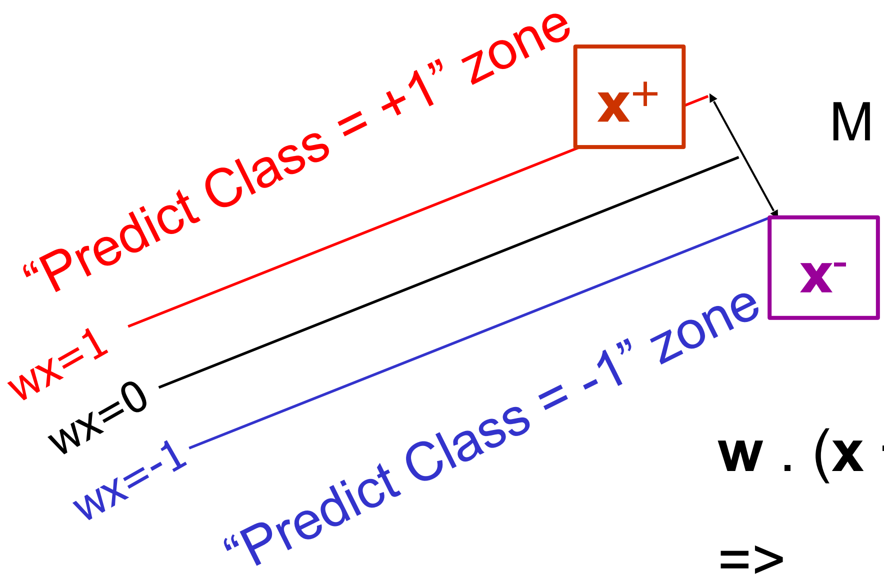
Computing the Margin Width



What we know:

- $\mathbf{w} \cdot \mathbf{x}^+ = +1$
- $\mathbf{w} \cdot \mathbf{x}^- = -1$
- $\mathbf{x}^+ = \mathbf{x}^- + \lambda \, \mathbf{w}$
- $|\mathbf{x}^+ - \mathbf{x}^-| = M$

Computing the Margin Width



M = Margin Width

What we know:

- $\mathbf{w} \cdot \mathbf{x}^+ = +1$
- $\mathbf{w} \cdot \mathbf{x}^- = -1$
- $\mathbf{x}^+ = \mathbf{x}^- + \lambda\, \mathbf{w}$
- $|\mathbf{x}^+ - \mathbf{x}^-| = M$

$$\mathbf{w} \cdot (\mathbf{x}^- + \lambda\, \mathbf{w}) + b = 1$$

=>

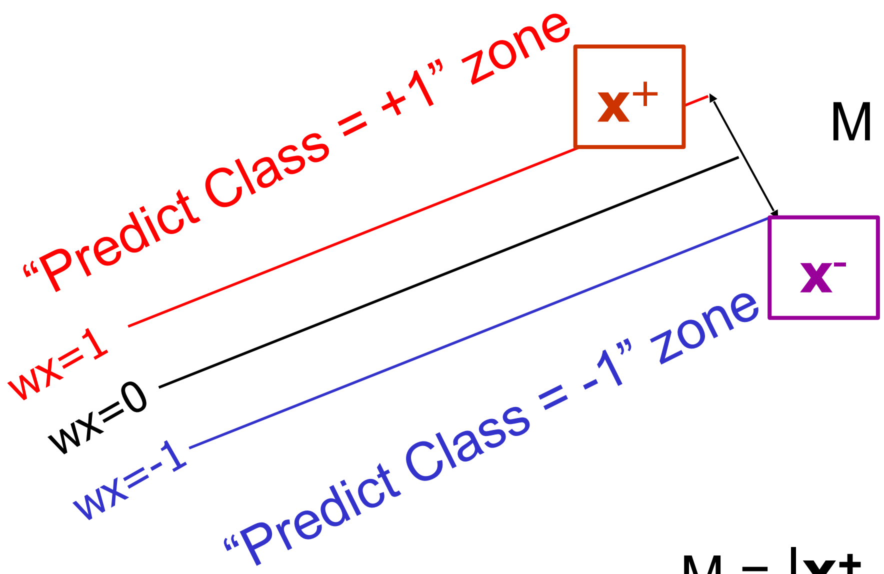
$$\mathbf{w} \cdot \mathbf{x}^- + b + \lambda\, \mathbf{w} \cdot \mathbf{w} = 1$$

=>

$$-1 + \lambda\, \mathbf{w} \cdot \mathbf{w} = 1$$

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

Computing the Margin Width



M = Margin Width

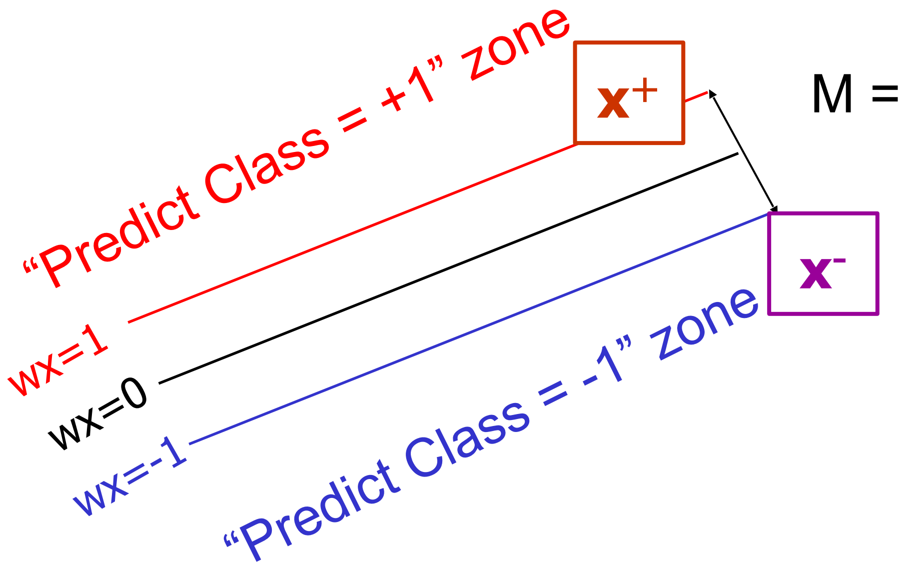
$$M = |\mathbf{x}^+ - \mathbf{x}^-| = |\lambda \mathbf{w}| = \lambda |\mathbf{w}| = \lambda \sqrt{\mathbf{w}.\mathbf{w}}$$

$$= \frac{2\sqrt{\mathbf{w}.\mathbf{w}}}{\mathbf{w}.\mathbf{w}} = \frac{2}{\sqrt{\mathbf{w}.\mathbf{w}}}$$

What we know:

- $\mathbf{w} \,.\, \mathbf{x}^+ = +1$
- $\mathbf{w} \,.\, \mathbf{x}^- = -1$
- $\mathbf{x}^+ = \mathbf{x}^- + \lambda \mathbf{w}$
- $|\mathbf{x}^+ - \mathbf{x}^-| = M$ $\quad \lambda = \dfrac{2}{\mathbf{w}.\mathbf{w}}$

Learning the Maximum Margin Classifier

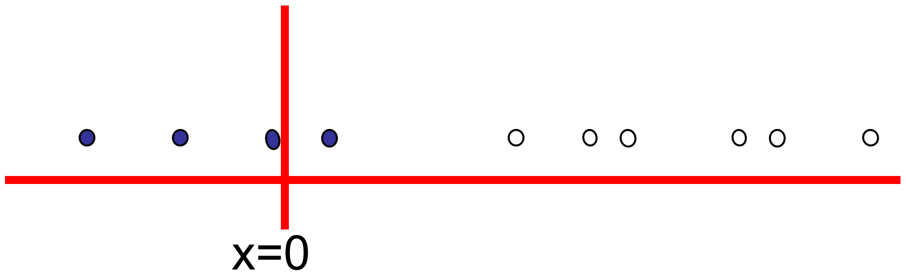


$$M = \text{Margin Width} = \frac{2}{\sqrt{\mathbf{w}.\mathbf{w}}}$$

Use optimization to search the space of W to find the widest margin that matches all the data points.

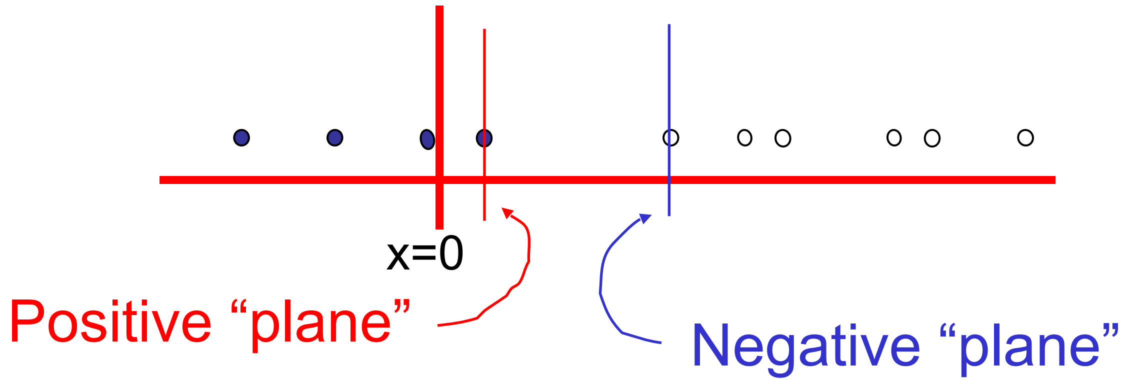
Simple 1-D Example

What would SVMs do?



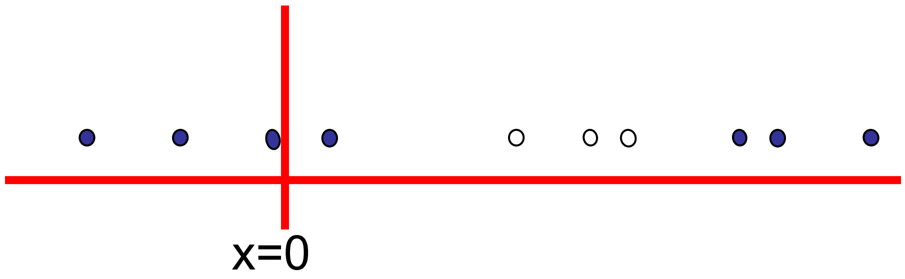
Simple 1-D Example

Not a big surprise

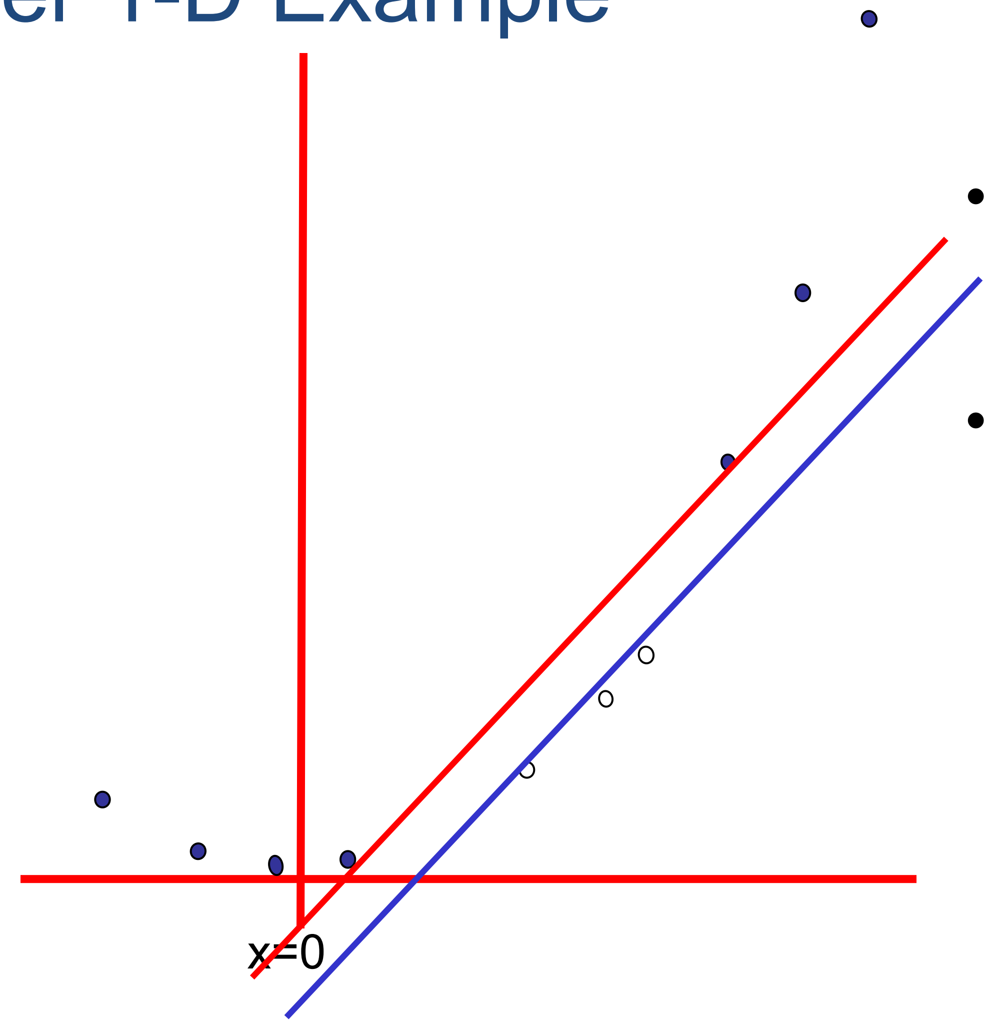


Harder 1-D Example

What can SVM do about this?



Harder 1-D Example

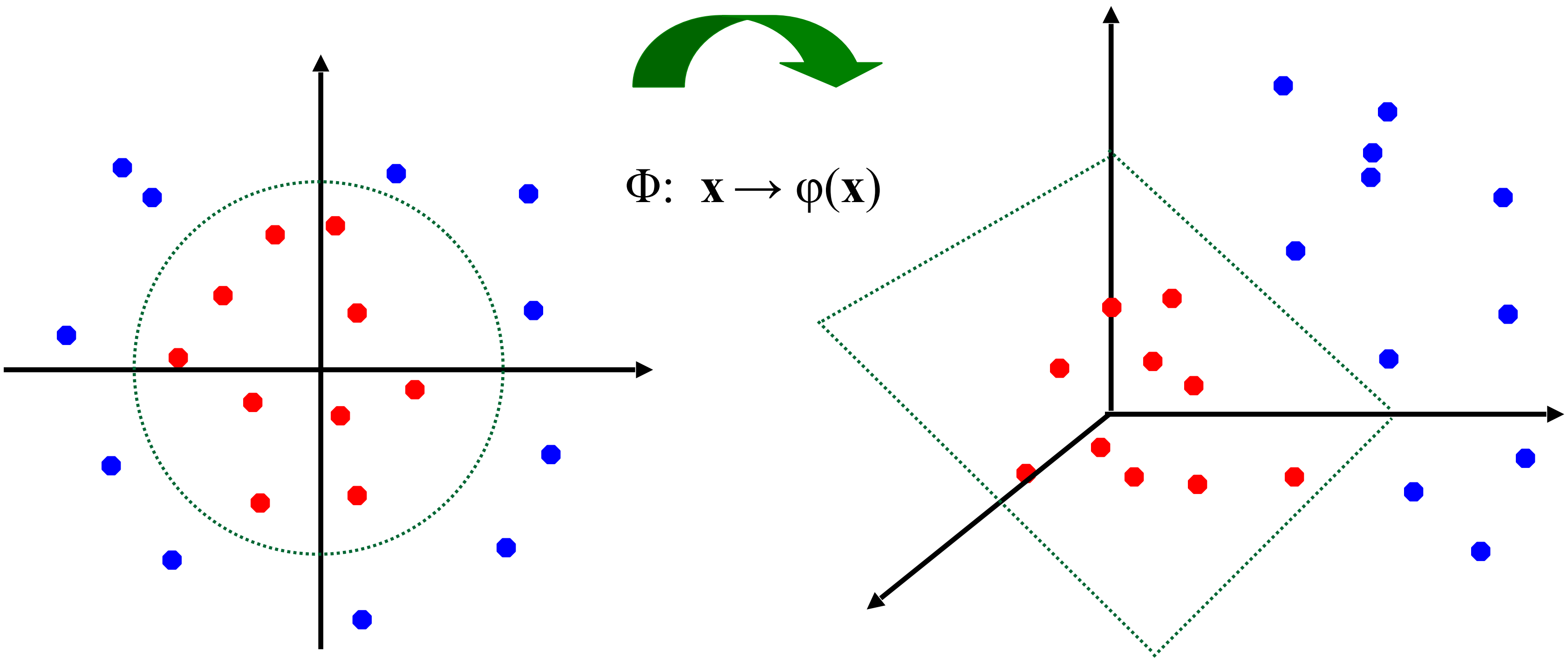


- Permit non-linear basis functions made linear regression
- Let's permit them here too

$$\mathbf{z}_k = (x_k, x_k^2)$$

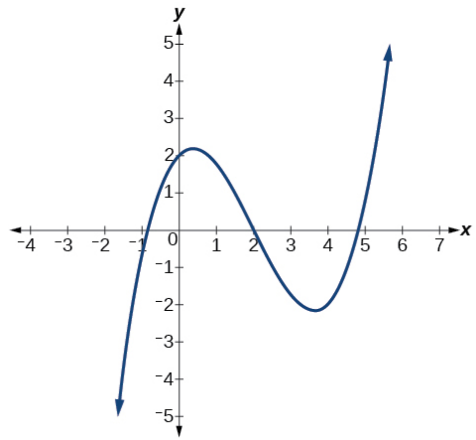
Nonlinear SVMs: Feature Space

- General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable:

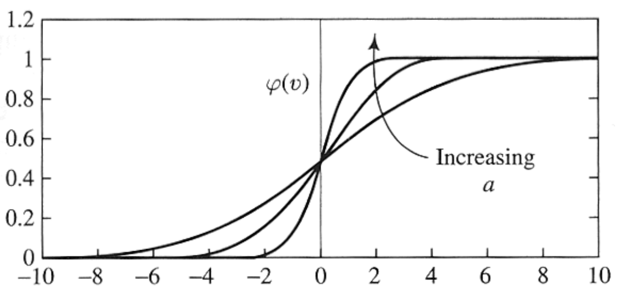


Nonlinear SVM Basis Functions

$\Phi(x_k)$ = (polynomial terms of $\mathbf{x}_k$ of degree 1 to q)



$\Phi(x_k)$ = (sigmoid functions of $\mathbf{x}_k$)



$\Phi(x_k)$ = (Gaussian radial basis functions of $\mathbf{x}_k$)