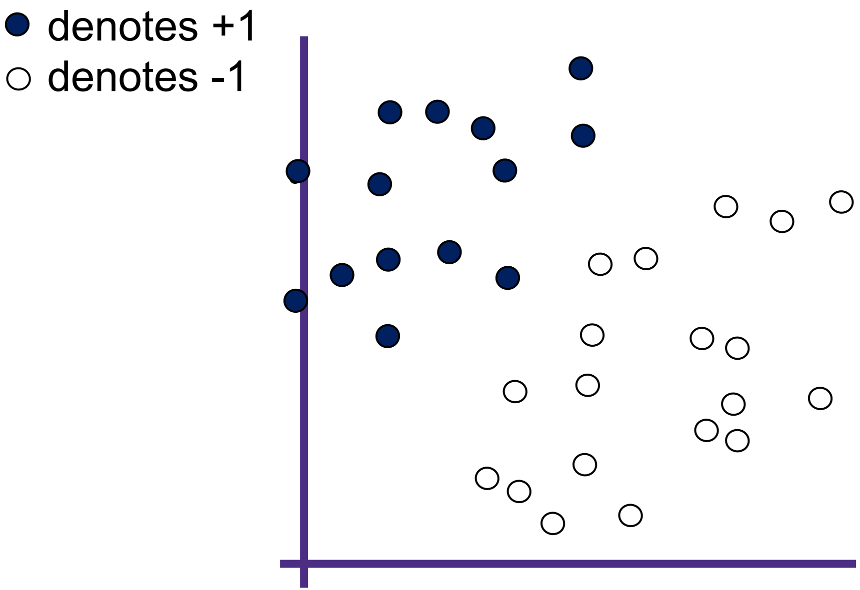
LECTURE 23

Support Vector Machine (SVM)

Maximal margin for classification and regression

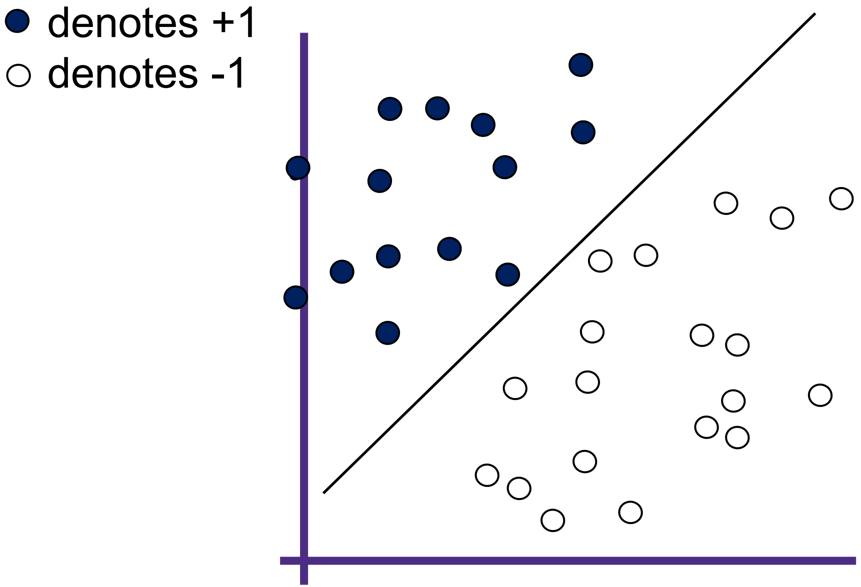
An example



Let X ∈ R^2, Y = {+1, -1}

How would you classify this data?

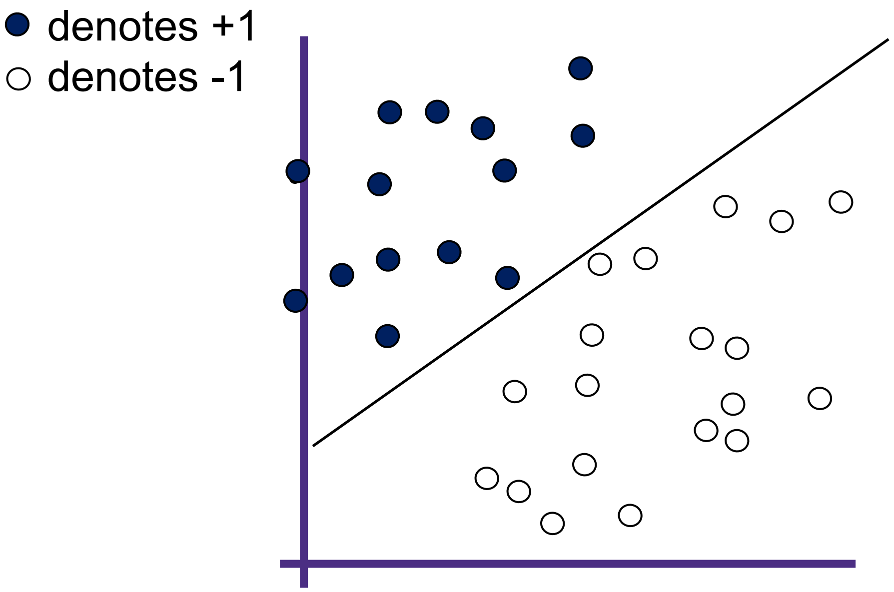
An example - Linear Classifiers



Let $X \in R^2$, $Y = \{+1, -1\}$

How would you classify this data?

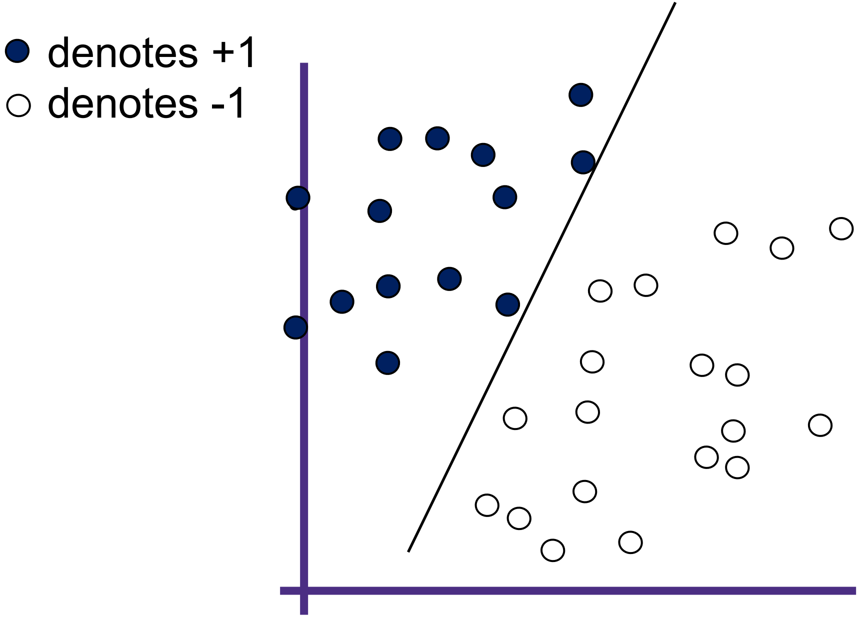
An example - Linear Classifiers



Let $X \in \mathbb{R}^2$, $Y = \{+1, -1\}$

How would you classify this data?

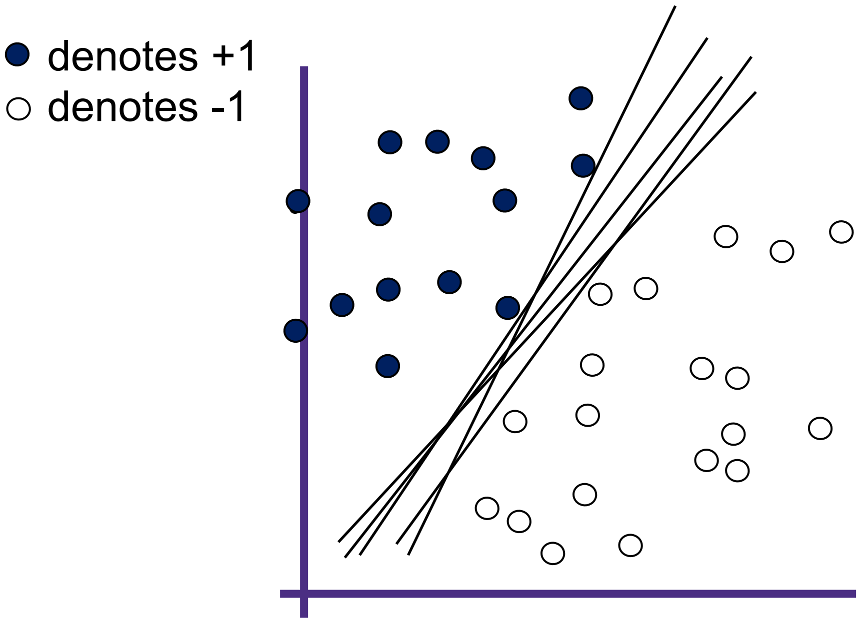
An example - Linear Classifiers



Let $X \in R^2$, $Y = \{+1, -1\}$

How would you classify this data?

An example - Linear Classifiers



Let X ∈ R^2, Y = {+1, -1}

How would you classify this data?

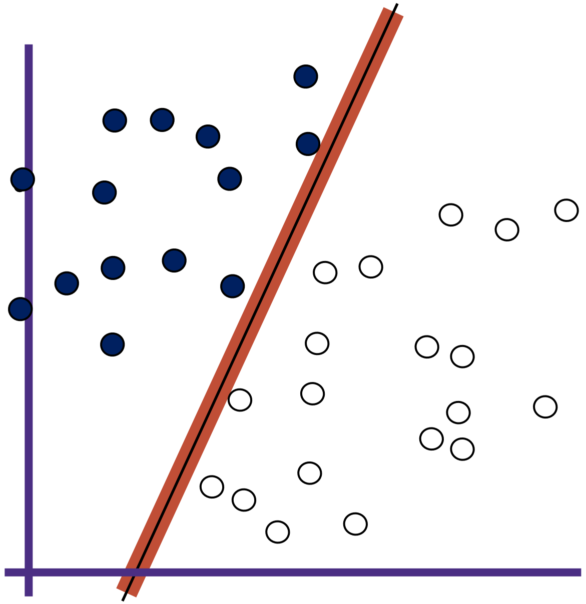
Any of these would be fine..

..but which is best?

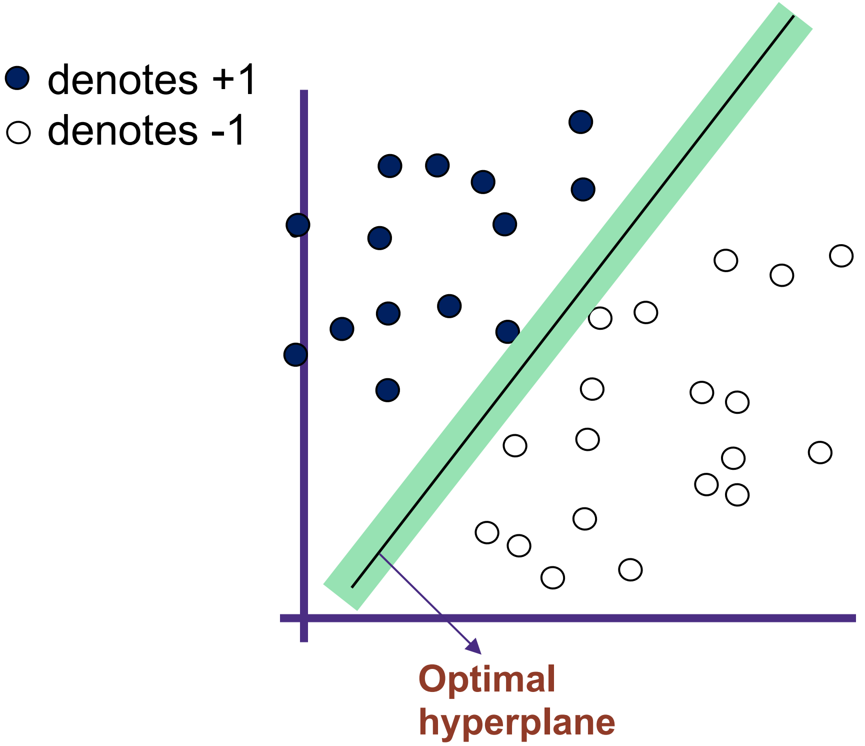
Classifier Margin

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

● denotes +1
○ denotes -1



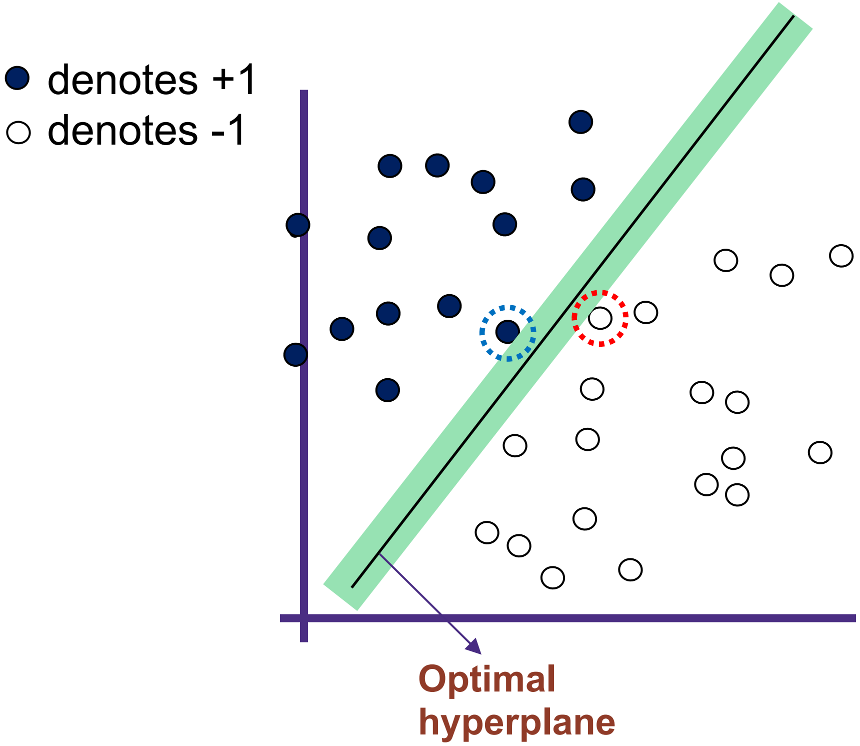
Maximum Margin



The **maximum margin linear classifier** is the linear classifier with the maximum margin.

This is the simplest kind of (linear) support vector machine.

Support Vector



Support vectors are data points that are closest to the optimal hyperplane.

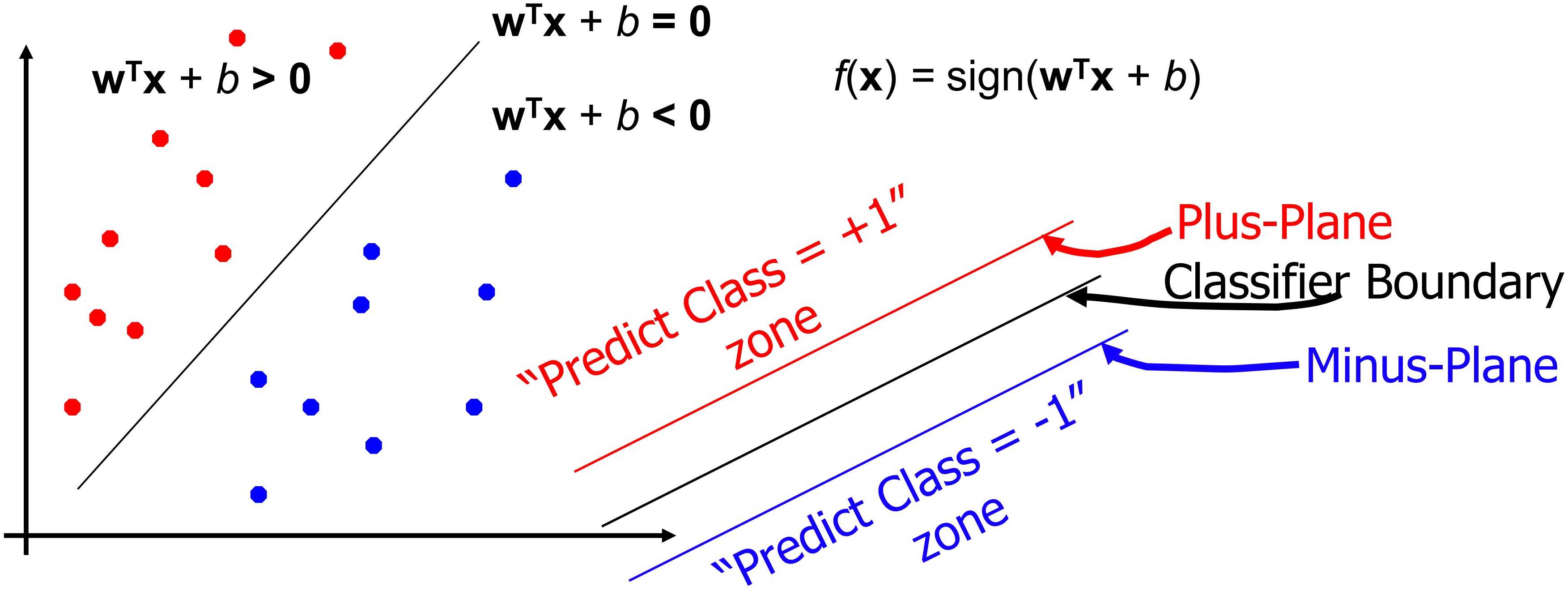
Implies that only support vectors matter; other training examples are ignorable.

Why maximum margin?

Maximizing the margin distance provides some reinforcement so that future data points can be classified with more confidence.`

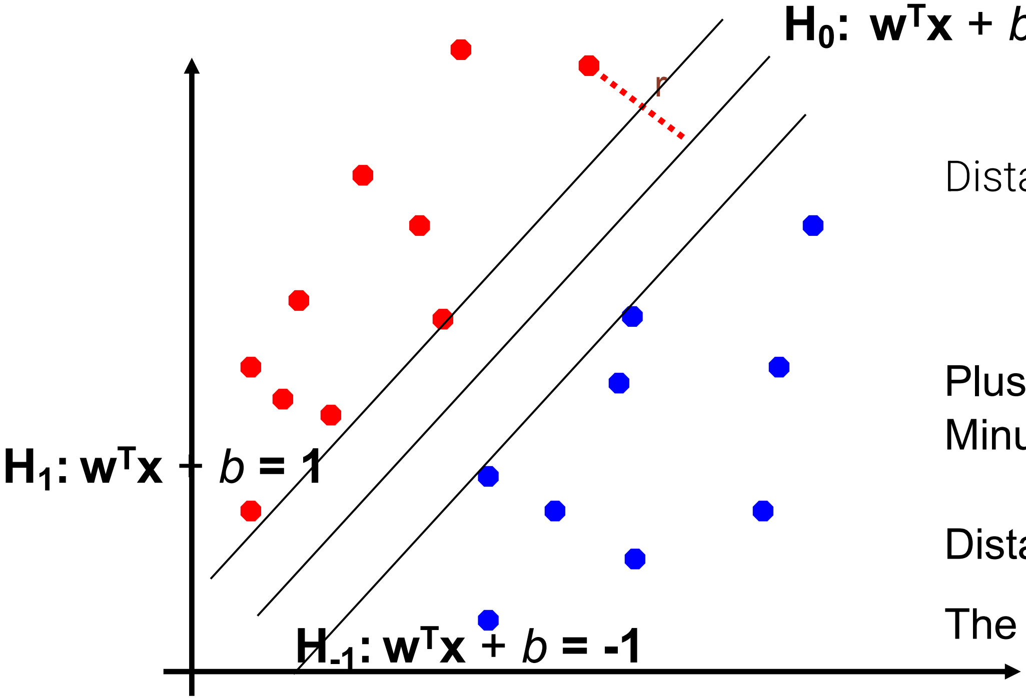
The model

- Binary classification can be viewed as the task of separating classes in feature space



The model

- Binary classification can be viewed as the task of separating classes in feature space



$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$$

Distance from example x_i to the hyperplane is

$$r = \frac{\mathbf{w}^T\mathbf{x}_i + b}{\|\mathbf{w}\|}$$

Plus-plane = $\{\, \boldsymbol{x} : \mathbf{w}^T\mathbf{x} + b = +1 \,\}$

Minus-plane = $\{\, \boldsymbol{x} : \mathbf{w}^T\mathbf{x} + b = -1 \,\}$

Distance from H_1 to H_0 is $\frac{1}{||w||}$

The margin width between H_1 and H_{-1} is $\frac{2}{||w||}$

Margin width

- M = Margin Width = $\frac{2}{||w||} = \frac{2}{\sqrt{w^T w}}$
- Claim: The vector **w** is perpendicular to the Plus Plane.

> Let u and v be two vectors on the Plus Plane.
> $w^T (u - v) = 0$

- Claim: the vector **w** is also perpendicular to the Minus Plane

Given a guess of ***w*** and *b* we can

- Compute whether all data points in the correct half-planes
- Compute the width of the margin

So now we just need to search for widest margin that matches all the datapoints.

Optimization: Quadratic Programming (QP)

Find **w** and *b* such that

$M = \frac{2}{||w||}$ is maximized

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1$

- QP is a well-studied class of optimization algorithms to maximize a quadratic function of some real-valued variables subject to linear constraints.
- Thus we reformulate the above problem as

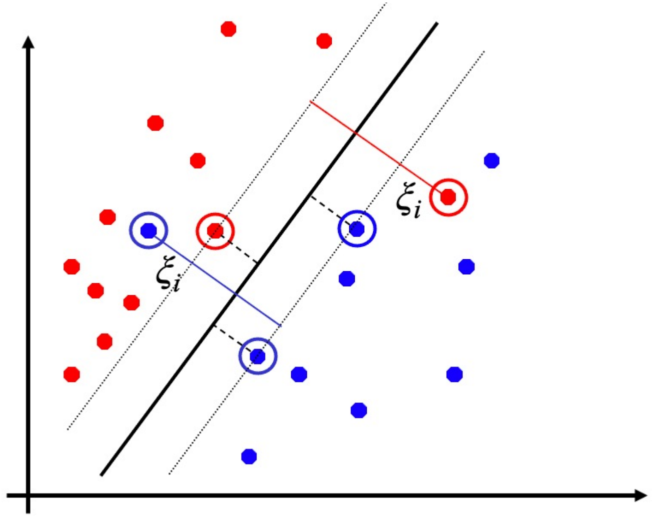
Find **w** and *b* such that

$\mathbf{\Phi}(\mathbf{w}) = ||\mathbf{w}||^2 = \mathbf{w}^\mathsf{T}\mathbf{w}$ is minimized

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^\mathsf{T}\mathbf{x}_i + b) \geq 1$

Hard and Soft Margin Classification

- If the data are “linearly separable”, then the algorithm is guaranteed to converge.
- What if the training set is not linearly separable?
  Slack variables ξ_i can be added to allow misclassification of difficult or noisy examples, resulting margin called *soft*.



Sometimes, the data is linearly separable, but the margin is so small that the model becomes prone to overfitting or being too sensitive to outliers. Also, in this case, we can opt for a larger margin by using soft margin SVM in order to help the model generalize better.

Soft Margin Classification

- The old formulation:

Find **w** and *b* such that

$\mathbf{\Phi}(\mathbf{w}) = \|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w}$ is minimized

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i (\mathbf{w}^T\mathbf{x}_i + b) \geq 1$

- Modified formulation incorporates slack variables:

Find w and b such that

$\mathbf{\Phi(w) = w^T w + C\Sigma\xi i}$ is minimized

and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i (\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$, , $\xi_i \geq 0$

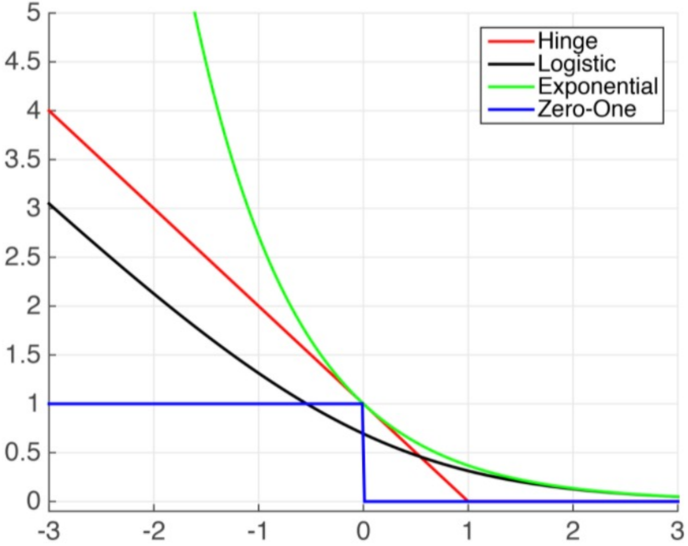
- Parameter *C* can be viewed as a way to control overfitting: it “trades off” the relative importance of maximizing the margin and fitting the training data.

Slack variables ξ_i

- From $y_i\,(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$, and $\xi_i \geq 0$
- We can conclude $\xi_i = \max(0, 1-y_i\,(\mathbf{w}^T\mathbf{x}_i + b))$

Actual Classification Loss

Hinge Loss $\max(0, 1 - y_i h(x_i))$	SVM
Log Loss $\log(1 + e^{-y_i \boldsymbol{h}(\boldsymbol{x}_i))})$	Logistic regression
Exponential Loss $e^{-y_i \boldsymbol{h}(\boldsymbol{x}_i))}$	AdaBoost
Zero-one Loss $\delta(\mathrm{sign}(\boldsymbol{h}(\boldsymbol{x}_i)) \neq \boldsymbol{y}_i)$	Actual Classification Loss (Not Impractical)

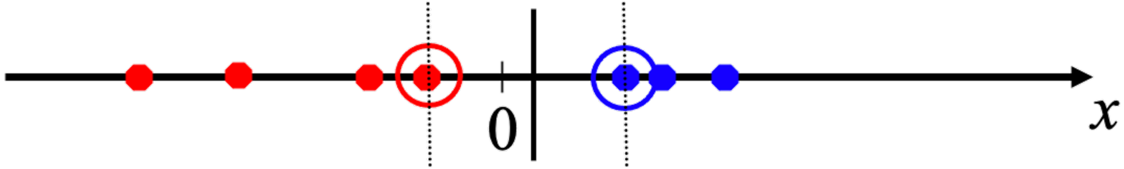


Summary – Linear SVM

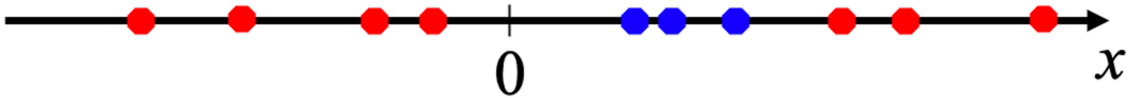
- The classifier is a *separating hyperplane.*
- Most "important" training points are support vectors; they define the hyperplane.
- Quadratic optimization algorithms can identify which training points $\mathbf{x}_i$ are support vectors with non-zero Lagrangian multipliers α_i.

Non-linear SVMs

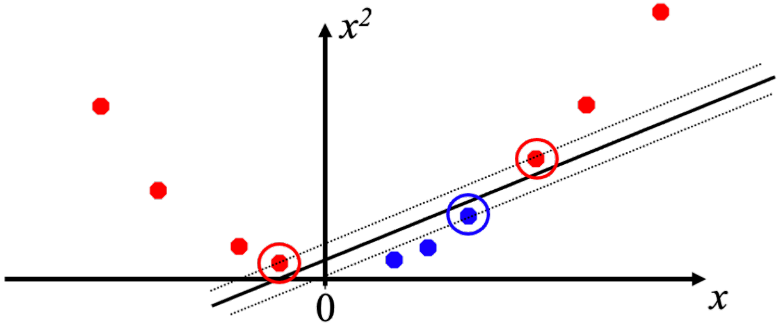
- Datasets that are linearly separable with some noise work out great:



- But what are we going to do if the dataset is just too hard?

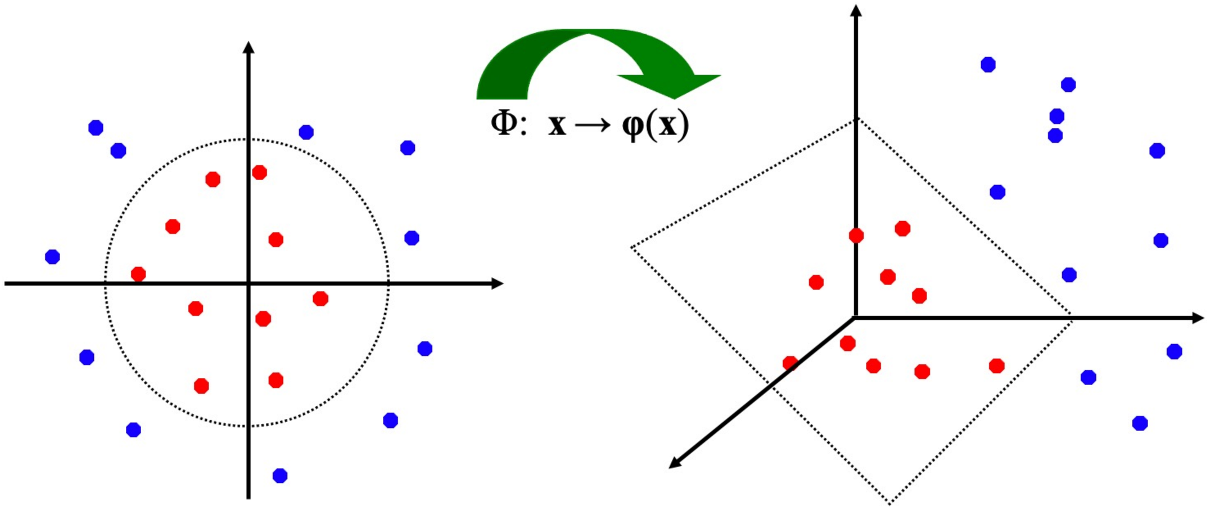


- How about… mapping data to a higher-dimensional space:

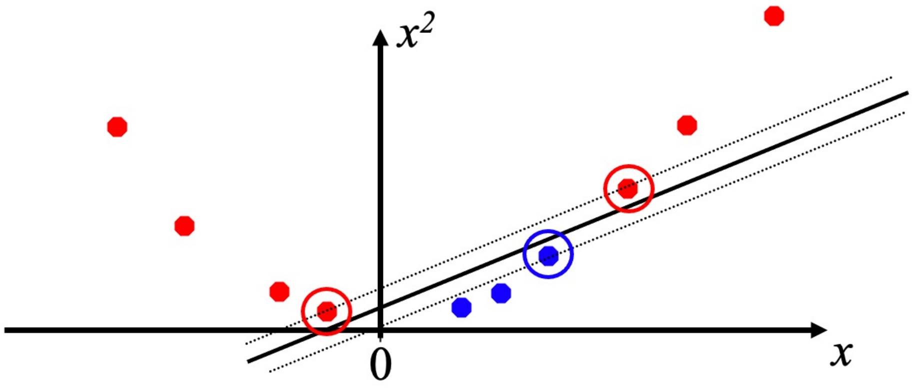


Non-linear SVMs: Feature spaces

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



Examples of Non-linear basis



$$Z = \varphi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$$

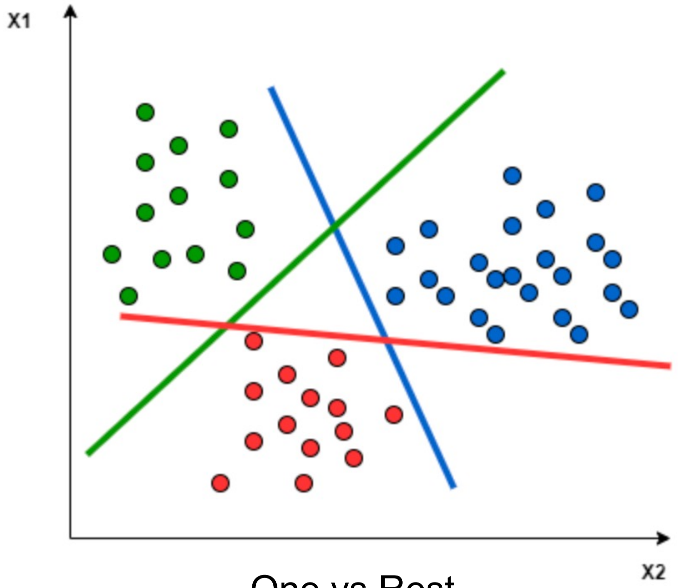
Find **w** and b such that

$\mathbf{\Phi}(\mathbf{w}) = \|\mathbf{w}\|^2 = \mathbf{w}^T\mathbf{w}$ is minimized

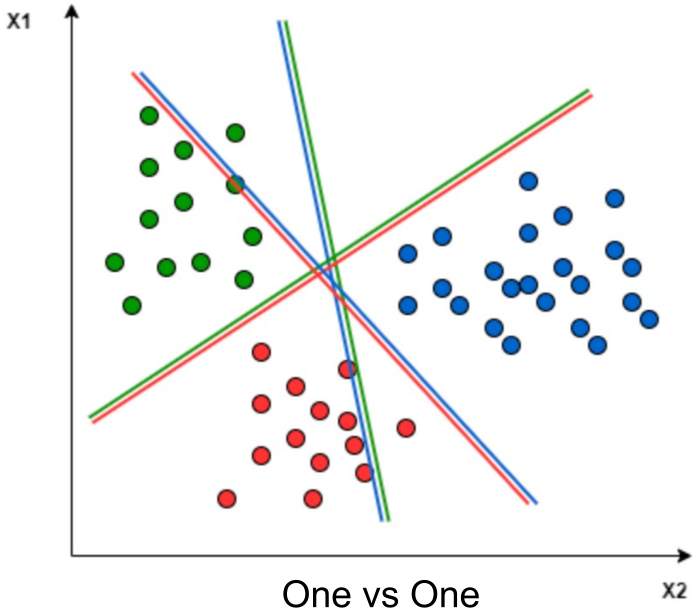
and for all $(\mathbf{z}_i, y_i)$, $i=1..n$:
$y_i\,(\mathbf{w}^T\mathbf{z}_i + b) \geq 1$

Multiclass SVM

•In the *One-to-Rest* approach, the classifier can use m SVMs. Each SVM would predict membership in one of the m classes.
•In the *One-to-One* approach, the classifier can use m(m-1)/2 SVMs.

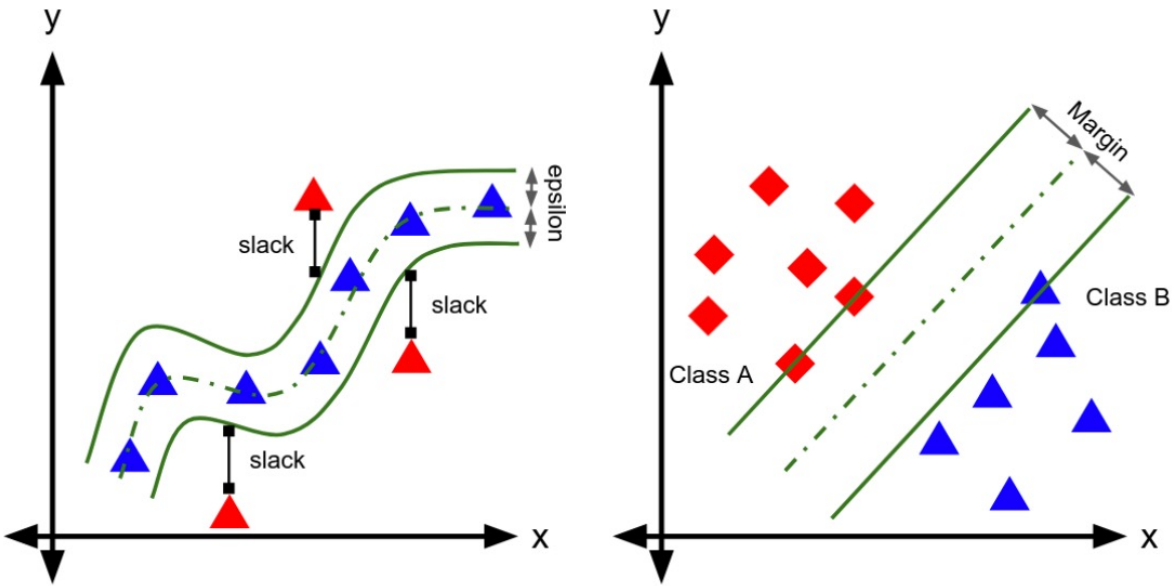


One vs Rest



One vs One

Support Vector Regression



Regression **Classification**

Take-home message

- Linear SVMs
- The definition of a maximum margin classifier
- What QP can do for you (but, for this class, you don’t need to know how it does it)
- How Maximum Margin can be turned into a QP problem
- How we deal with noisy (non-separable) data
- How we permit non-linear boundaries
- How we extend the algorithm for multi-class classification and regression