Support Vector Machines: Part 1

Outline

- pre: Linear Classifiers and Hyperplanes
- SVM overview: Geometric Interpretation
- Linear SVM: the Original
- Linear SVM: Soft Margins
- Tomorrow Part II: Non-linear SVMs (kernels)

Support Vector Machines

Linear classifier (classifiers with a linear decision boundary)

Support Vector Machines

Linear classifier (classifiers with a linear decision boundary)

Geometrically Motivated

Support Vector Machines

Linear classifier (classifiers with a linear decision boundary)

Geometrically Motivated

Originally proposed for binary classifications (two classes)

Support Vector Machines

Linear classifier (classifiers with a linear decision boundary)

Geometrically Motivated

Originally proposed for binary classifications (two classes)

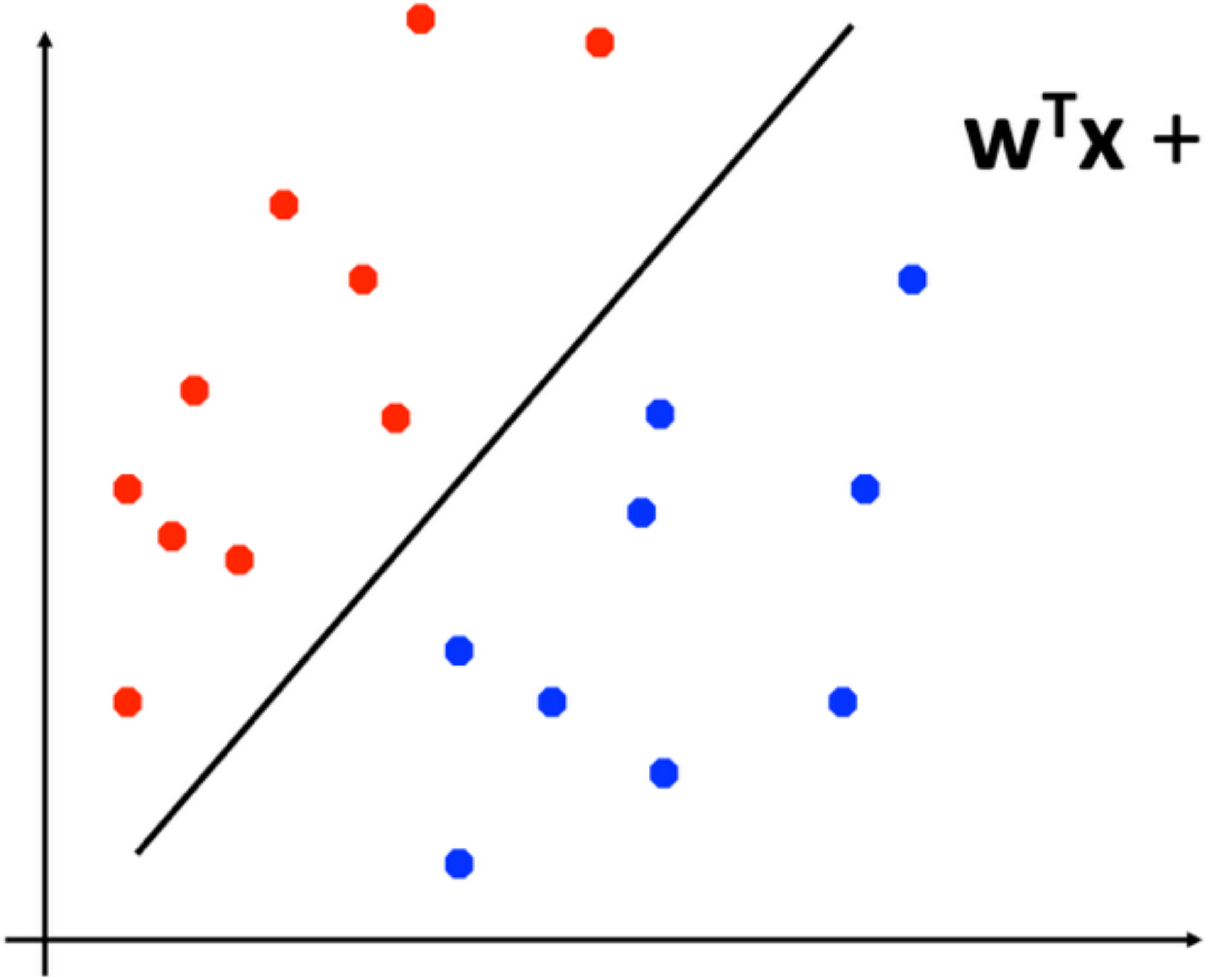
Has been extended to handle multiple-class classifications as well as regressions

Linear Decision Boundaries

$$w_1 x_1 + b = 0$$

$$\mathbf{w}^T\mathbf{x} + b = 0$$

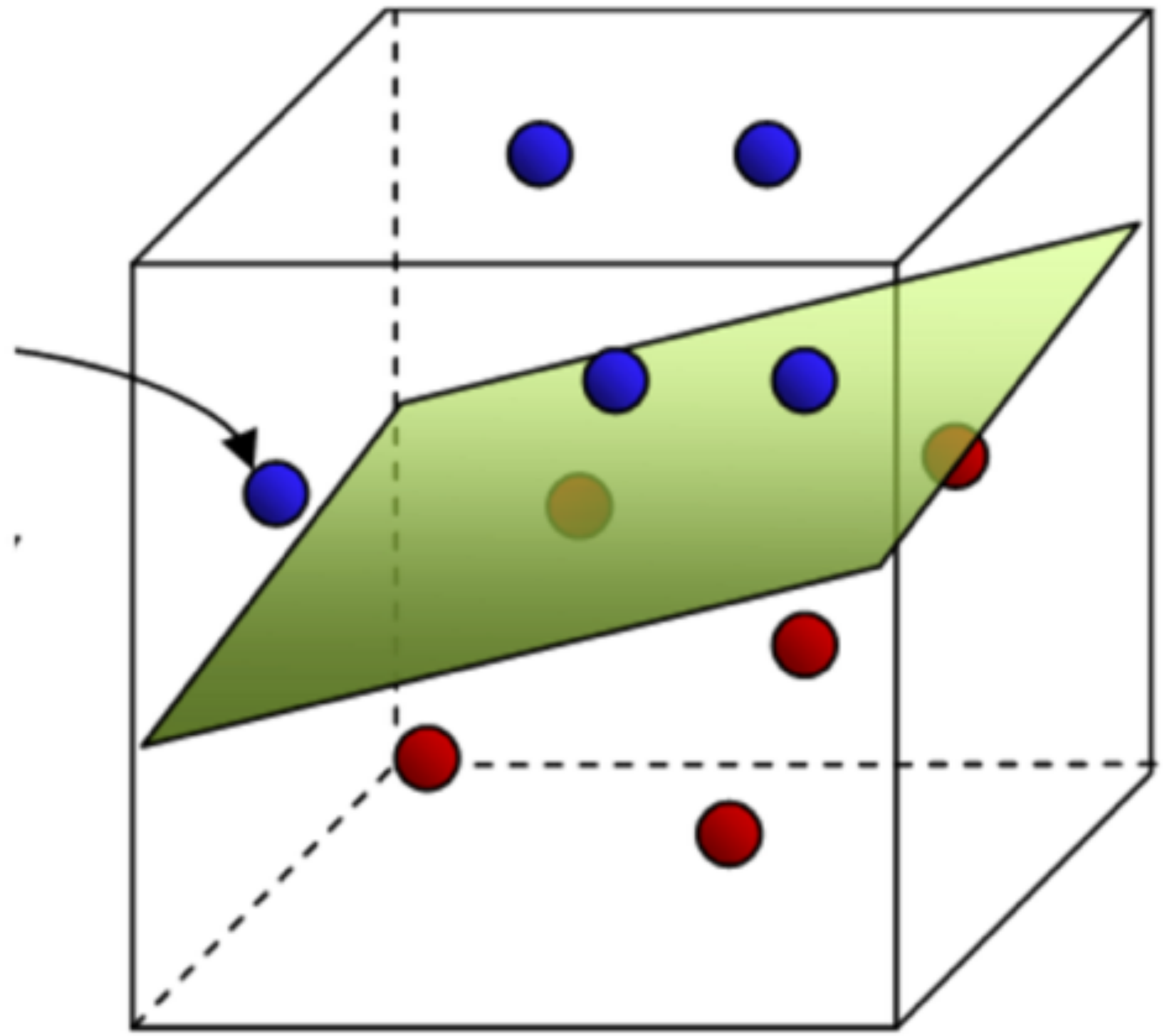
Linear Decision Boundaries



$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$

$$w_1 x_1 + w_2 x_2 + b = 0$$

Linear Decision Boundaries



$$\mathbf{w}^T\mathbf{x} + b = 0$$

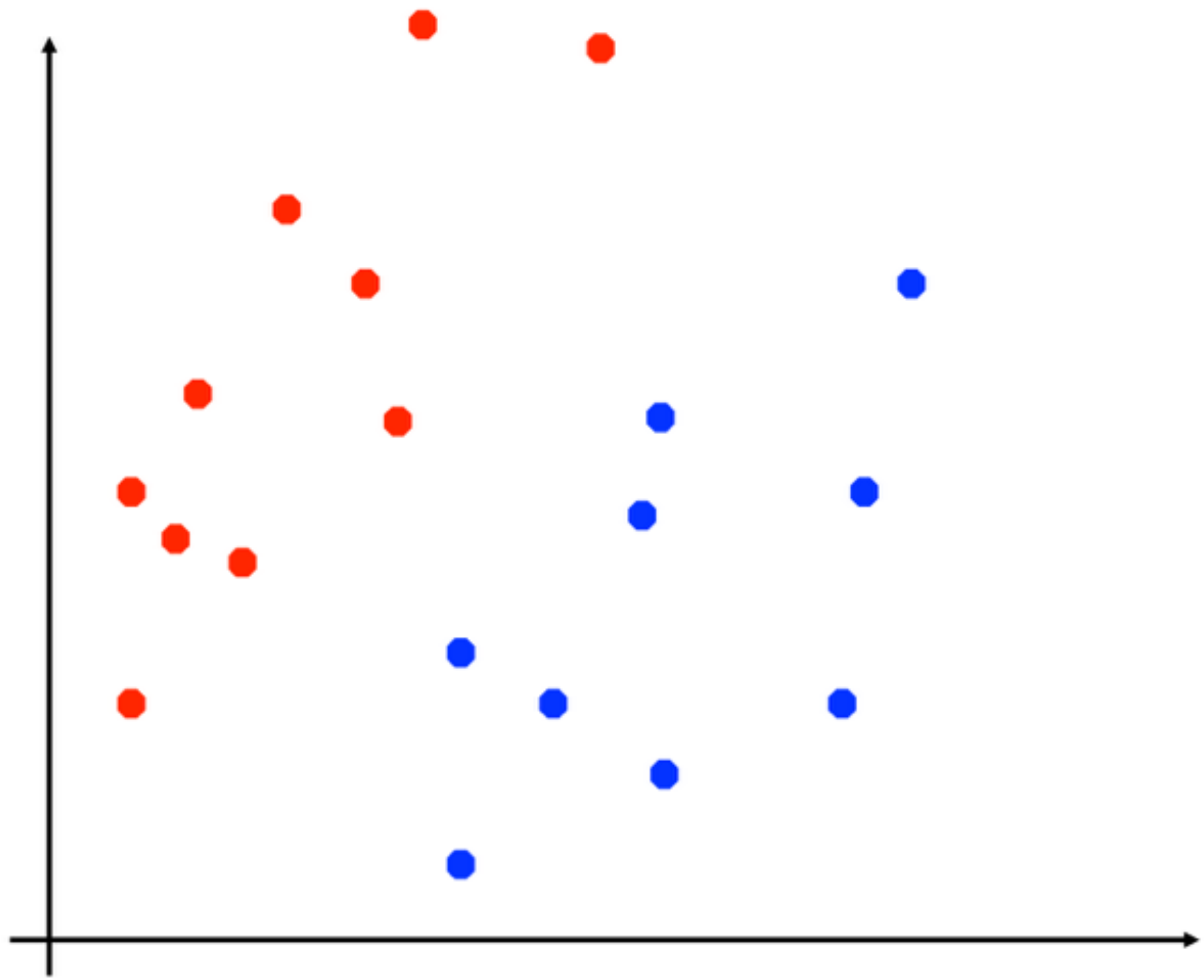
$$w_1x_1 + w_2x_2 + w_3x_3 + b = 0$$

Linear Decision Boundaries

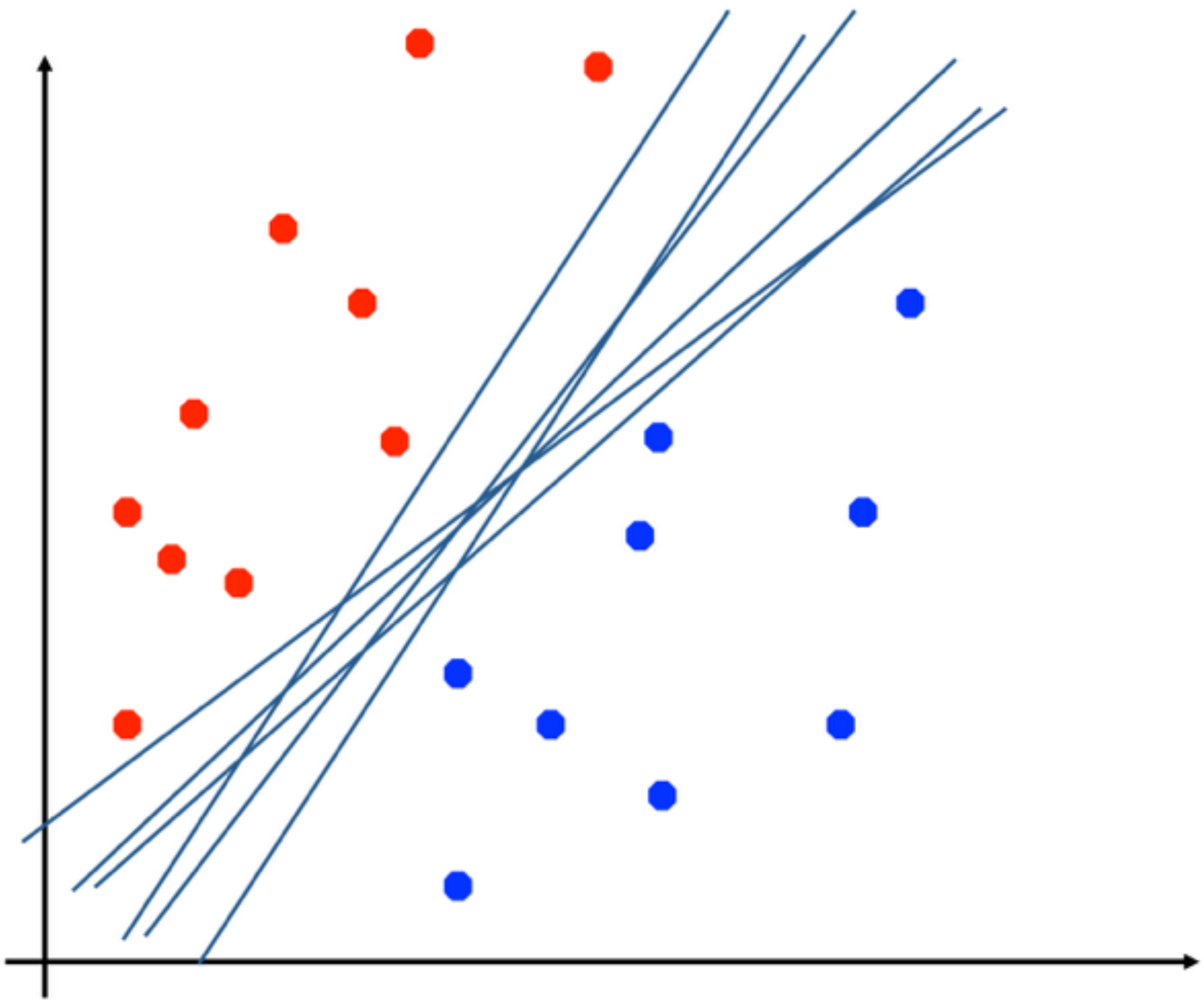
$$w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b = 0$$

$$\mathbf{w}^{\mathsf{T}}\mathbf{x} + b = 0$$

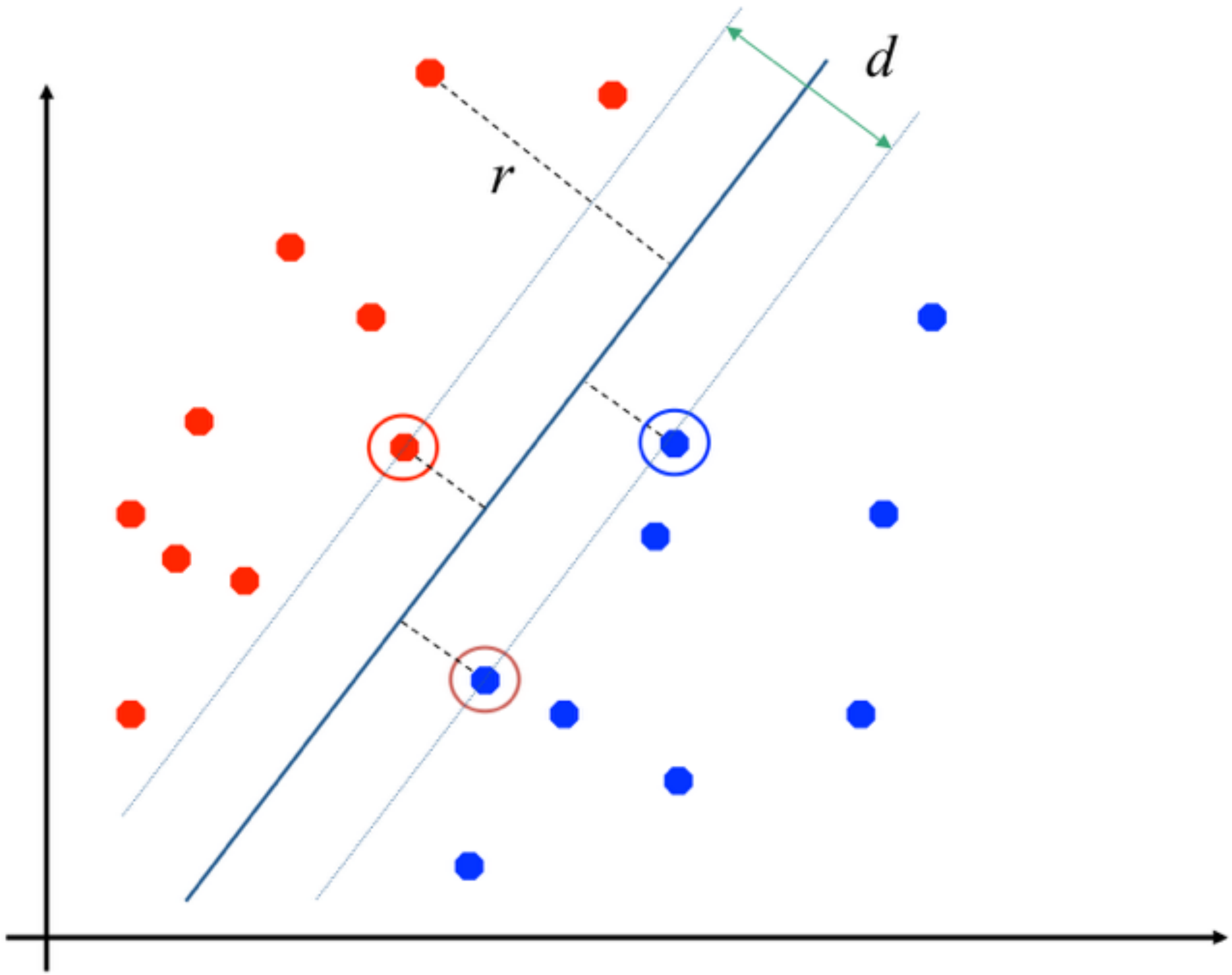
SVM: Geometric Interpretation



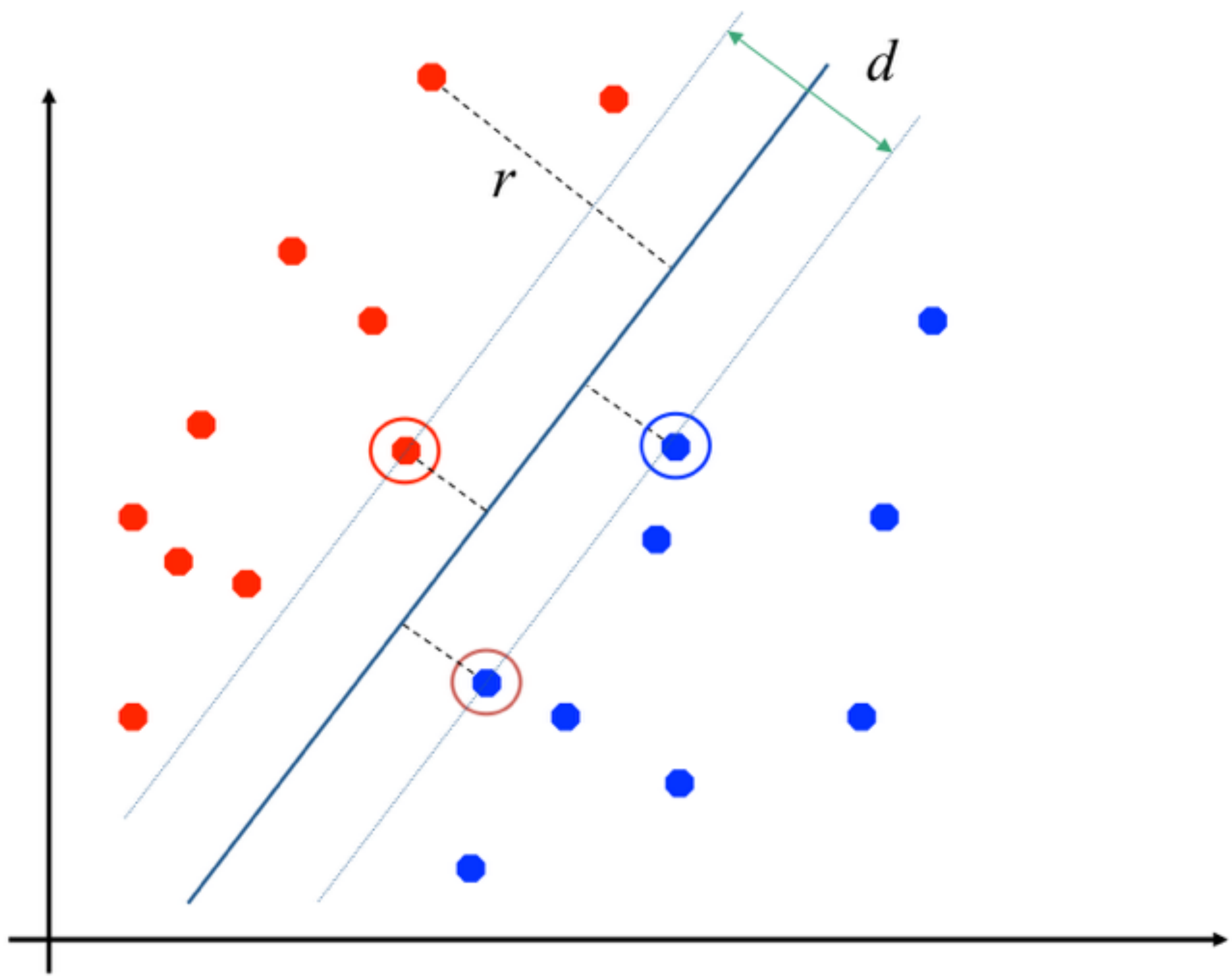
SVM: which is the optimal separator?



Maximizing the Margin



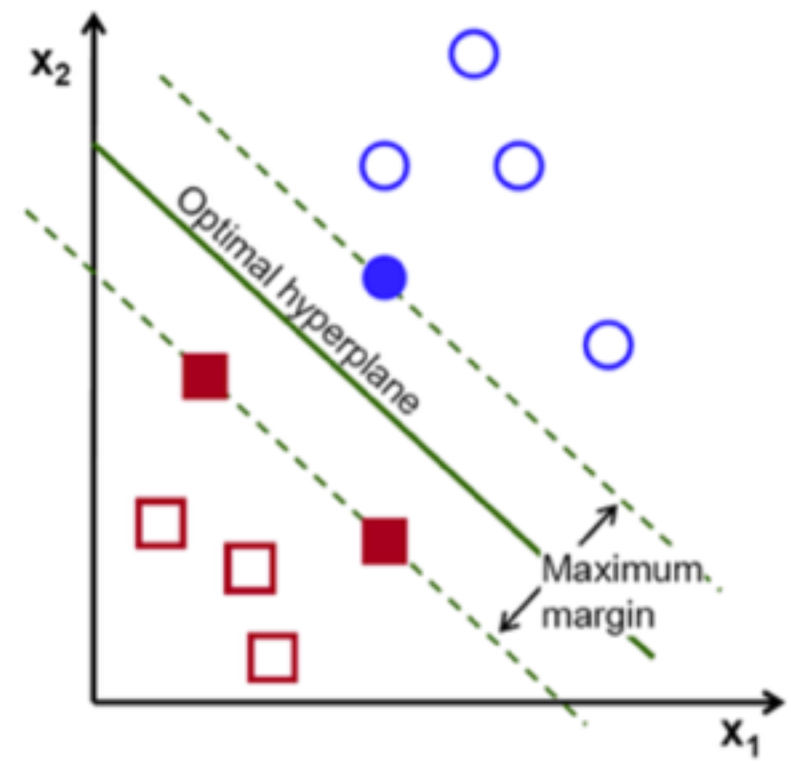
Maximizing the Margin



Interview Time: SVM Basics

What is SVM?

- linear classifier for binary classification



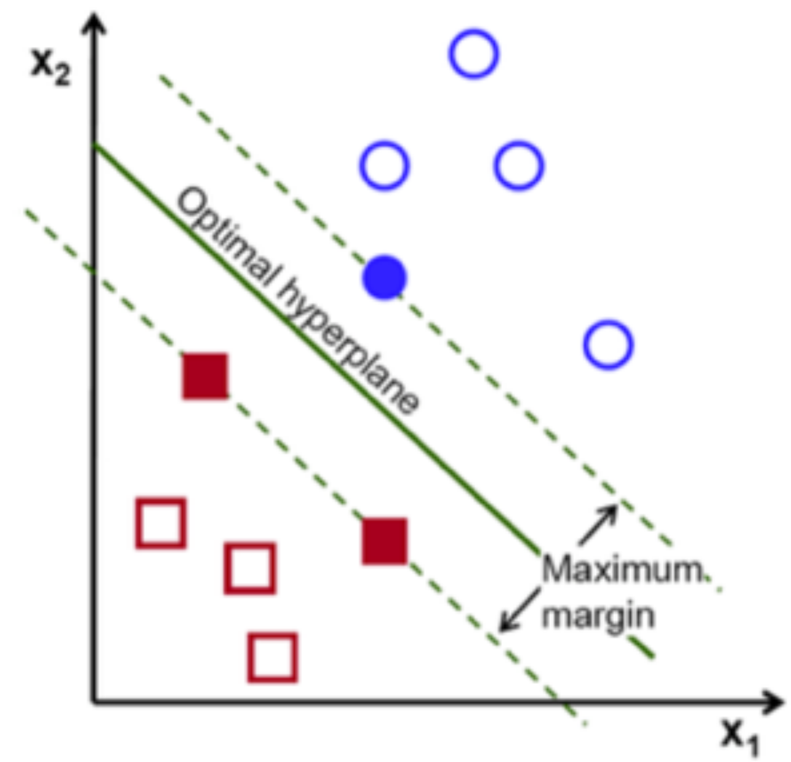
Interview Time: SVM Basics

What is SVM?

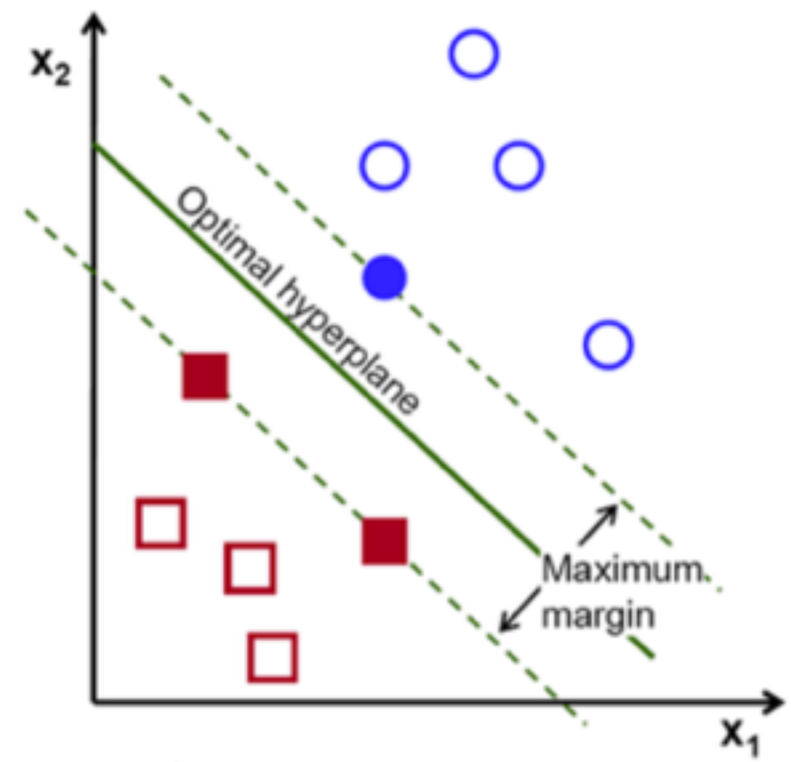
- linear classifier for binary classification

How does it work?

- finds optimal separating hyperplane that has the maximum margin



Interview Time: SVM Basics



What is SVM?

- linear classifier for binary classification

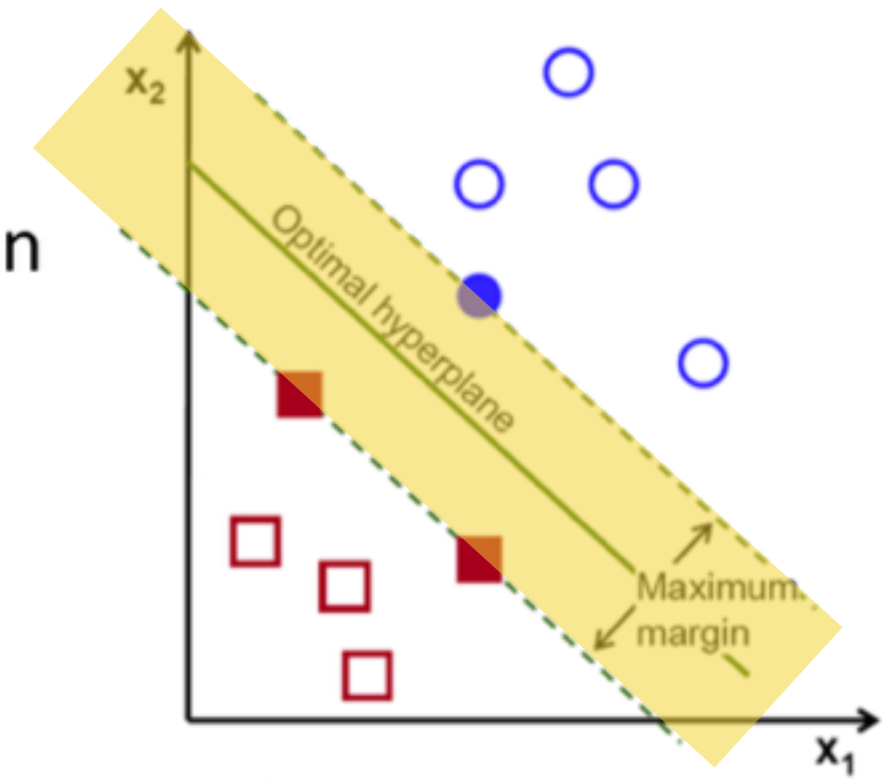
How does it work?

- finds optimal separating hyperplane that has the maximum margin

What do you mean margin?

- distance between the closest data points to the separator, margin is "no man's land", no data point inside margin

Interview Time: SVM Basics



What is SVM?

- linear classifier for binary classification

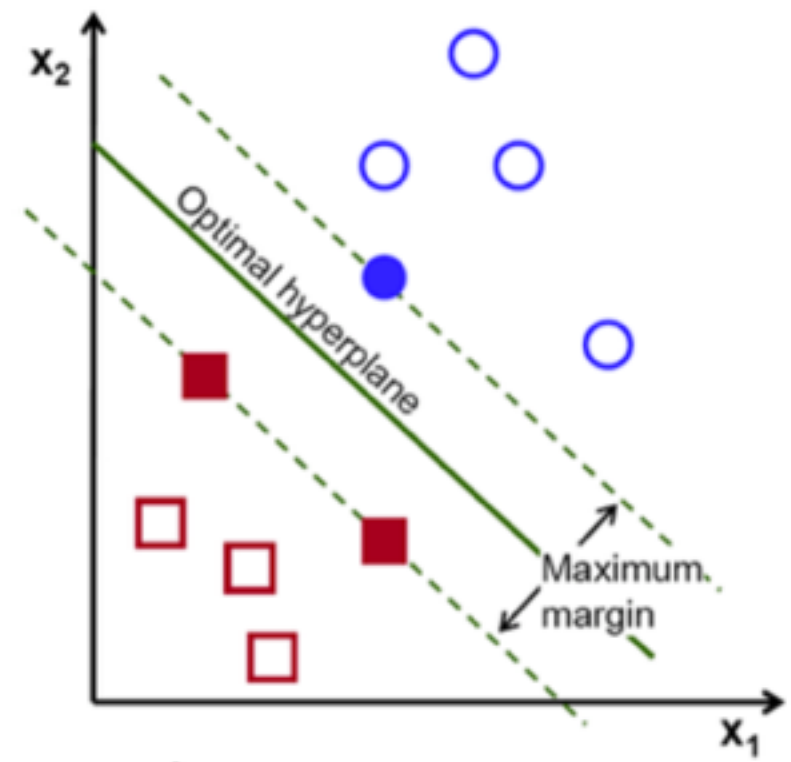
How does it work?

- finds optimal separating hyperplane that has the maximum margin

What do you mean margin?

- distance between the closest data points to the separator, margin is “no man’s land”, no data point inside margin

Interview Time: SVM Basics



What is SVM?

- linear classifier for binary classification

How does it work?

- finds optimal separating hyperplane that has the maximum margin

What do you mean margin?

- distance between the closest data points to the separator, margin is "no man's land", no data point inside margin

What are the support vectors?

- a subset of your data, those closest to the decision boundary, they define the hyperplane, rest data points essentially irrelevant

Interview Time: SVM Basics

How do you find that optimal boundary with the maximized margin?

Interview Time: SVM Basics

How do you find that optimal boundary with the maximized margin?

Interview Time: SVM Basics

How do you find that optimal boundary with the maximized margin?

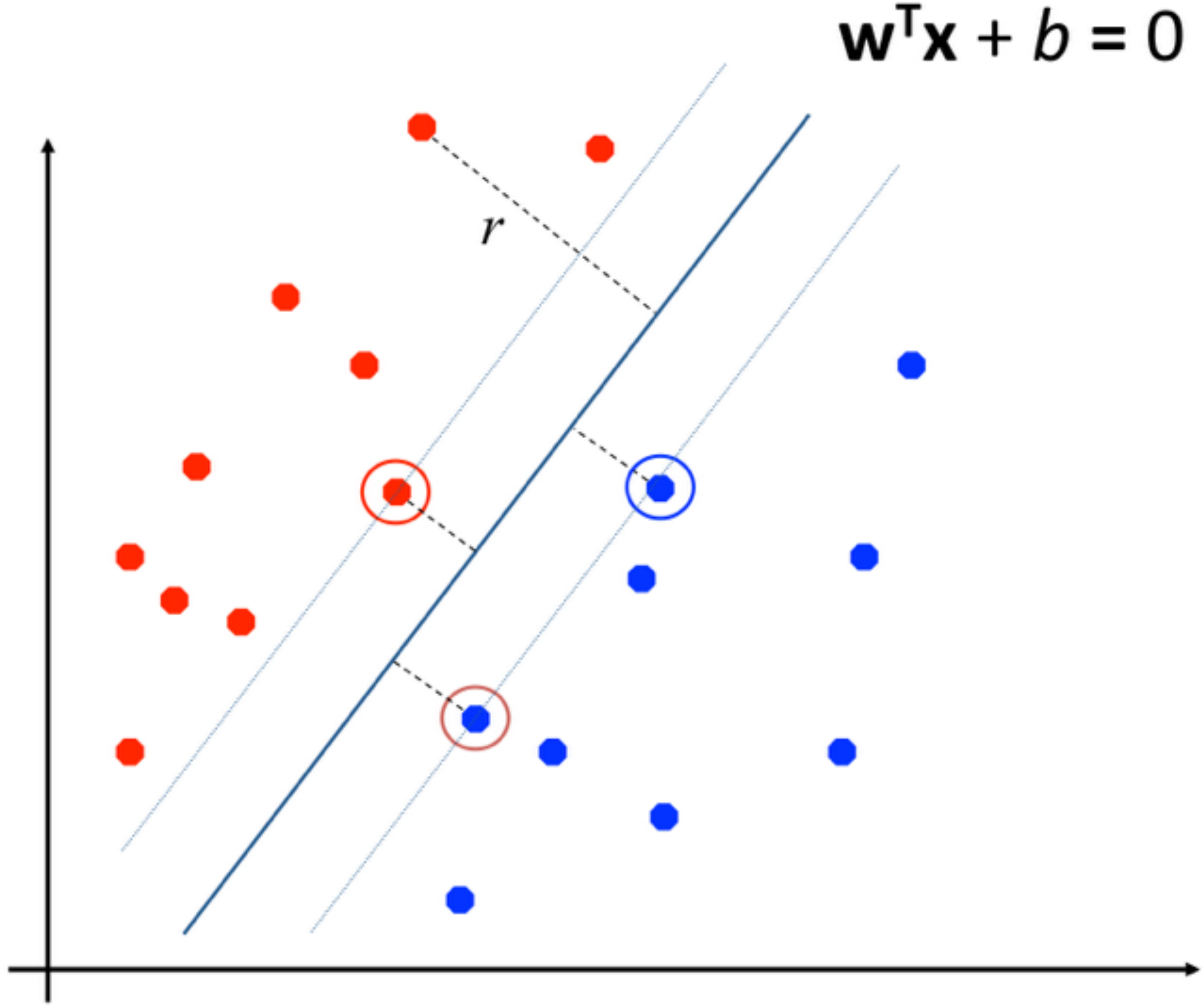
…by solving some sort of optimization problem mathematically?

Interview Time: SVM Basics

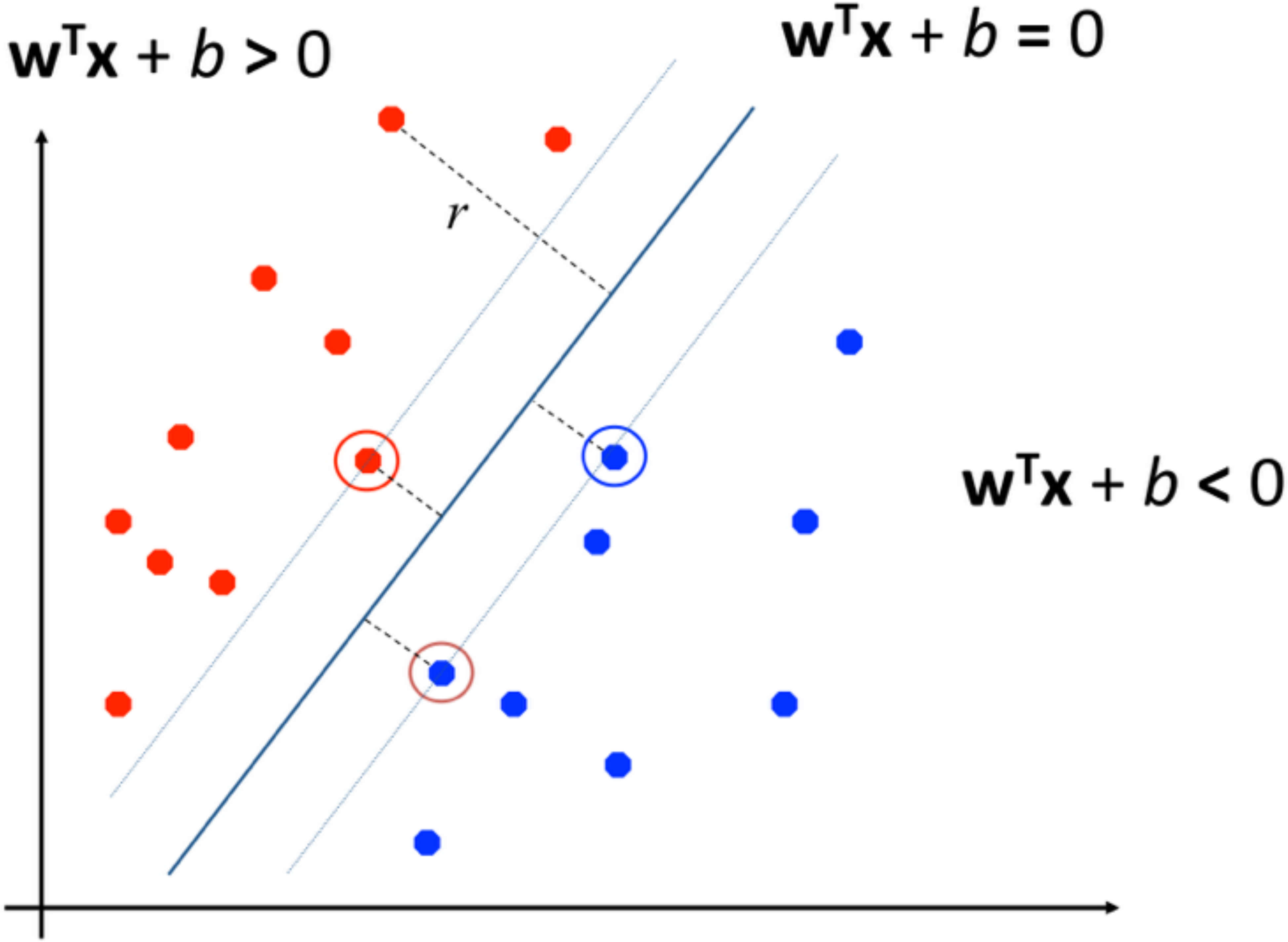
How do you find that optimal boundary with the maximized margin?

…by solving some sort of optimization problem mathematically?

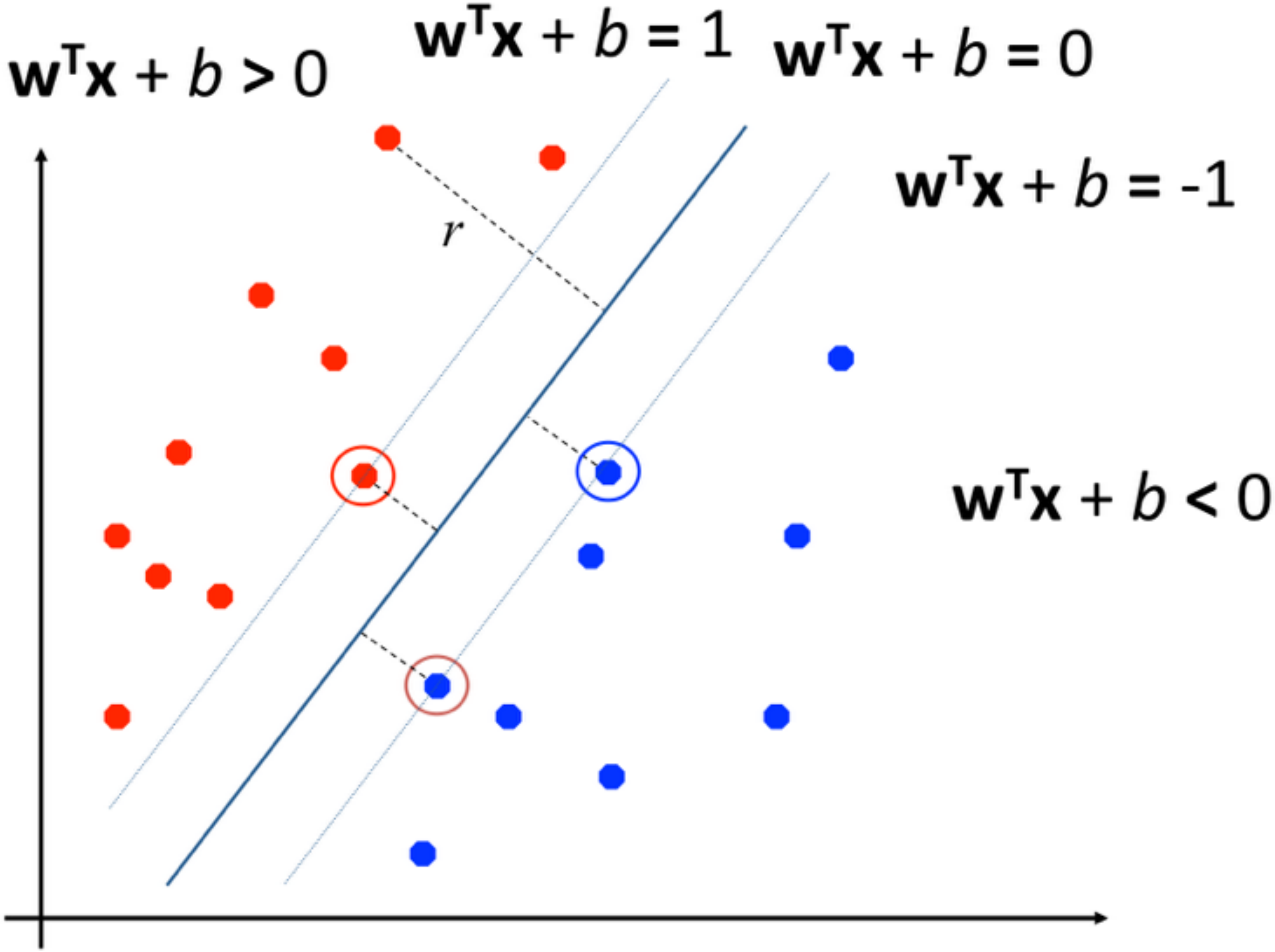
Linear SVMs Mathematically



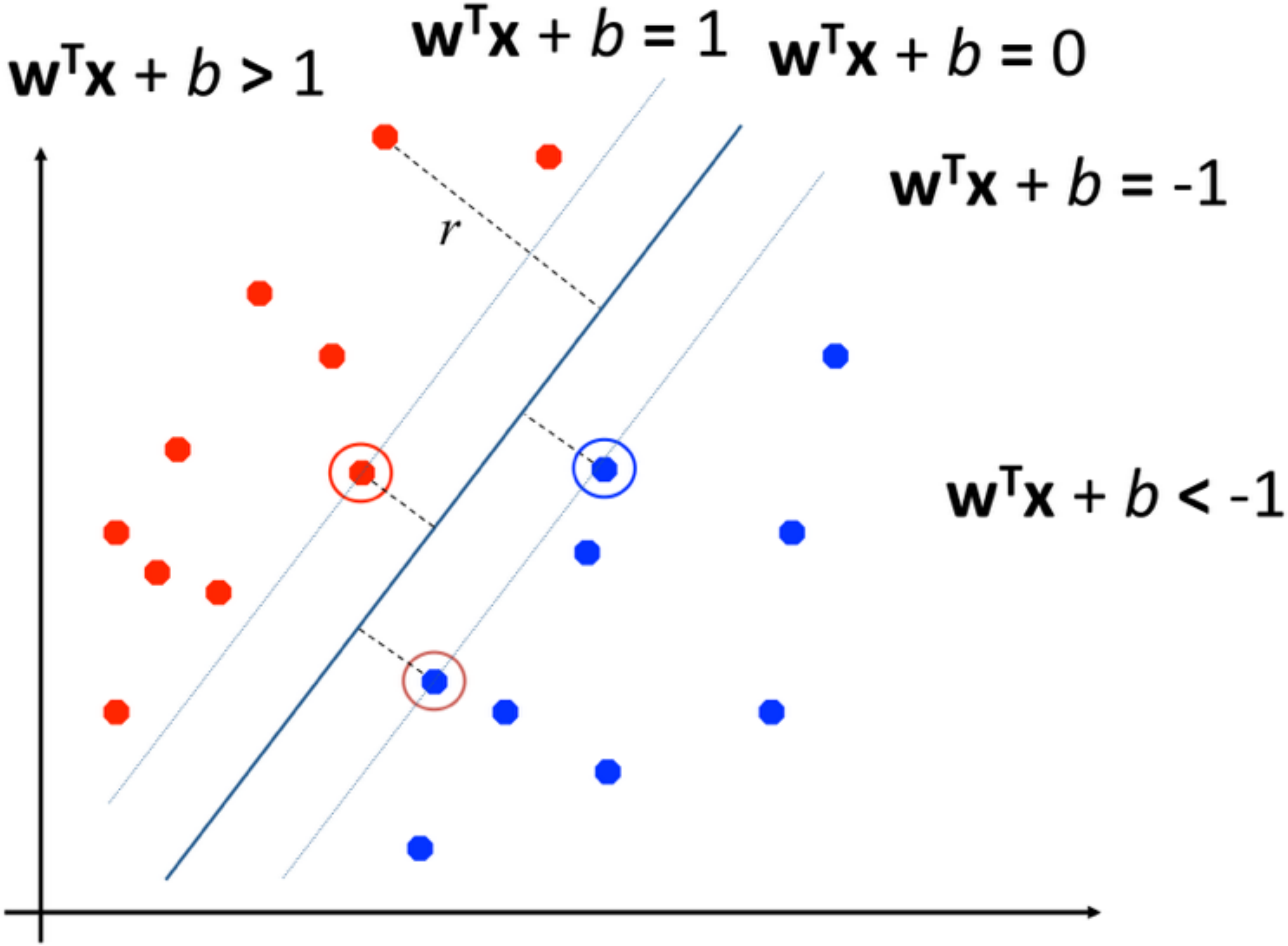
Linear SVMs Mathematically



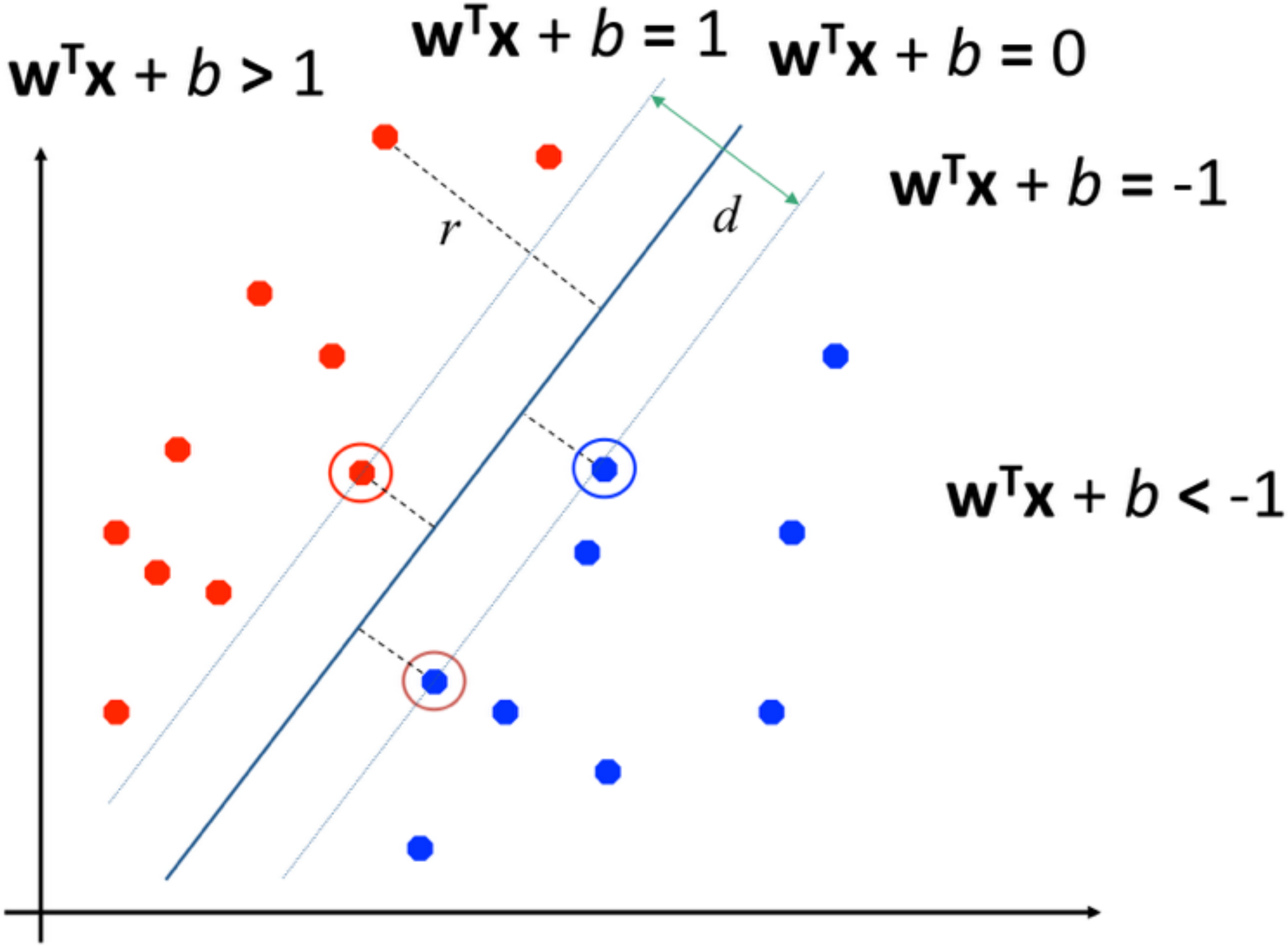
Linear SVMs Mathematically



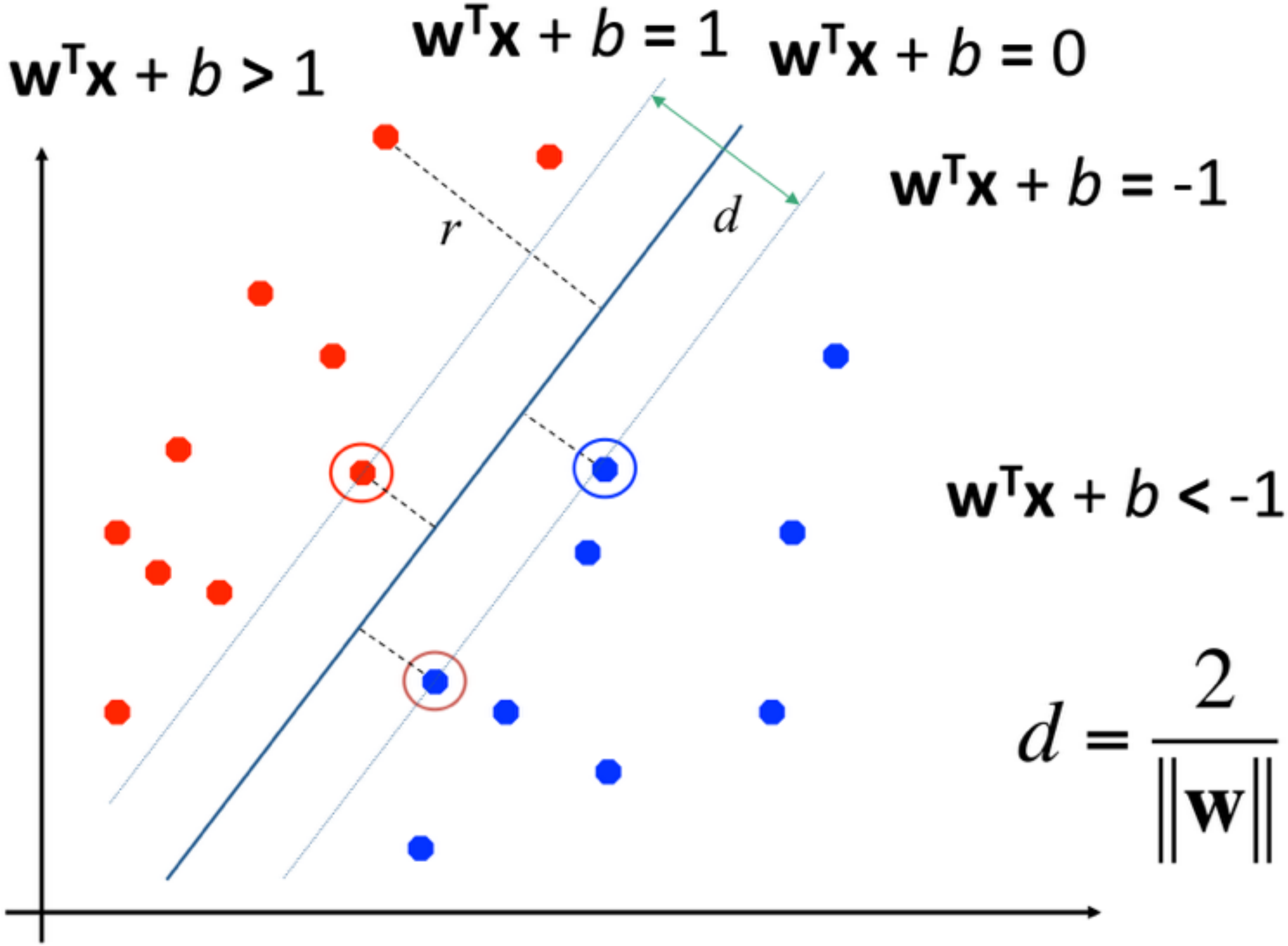
Linear SVMs Mathematically



Linear SVMs Mathematically



Linear SVMs Mathematically



Linear SVMs Mathematically (cont.)

Then we can formulate the *quadratic optimization problem:*

Find $\mathbf{w}$ and b such that

$$d = \frac{2}{\|\mathbf{w}\|} \text{ is maximized}$$

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

Which can be reformulated as:

Find $\mathbf{w}$ and b such that

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

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

The Optimization Problem Solution

- Given a solution $\alpha_1 \ldots \alpha_n$ to the dual problem, solution to the primal is:

$$\boxed{\mathbf{w} = \Sigma \alpha_i y_i \mathbf{x}_i \qquad b = y_k - \Sigma \alpha_i y_i \mathbf{x}_i^{\mathbf{T}} \mathbf{x}_k \quad \text{for any } \alpha_k > 0}$$

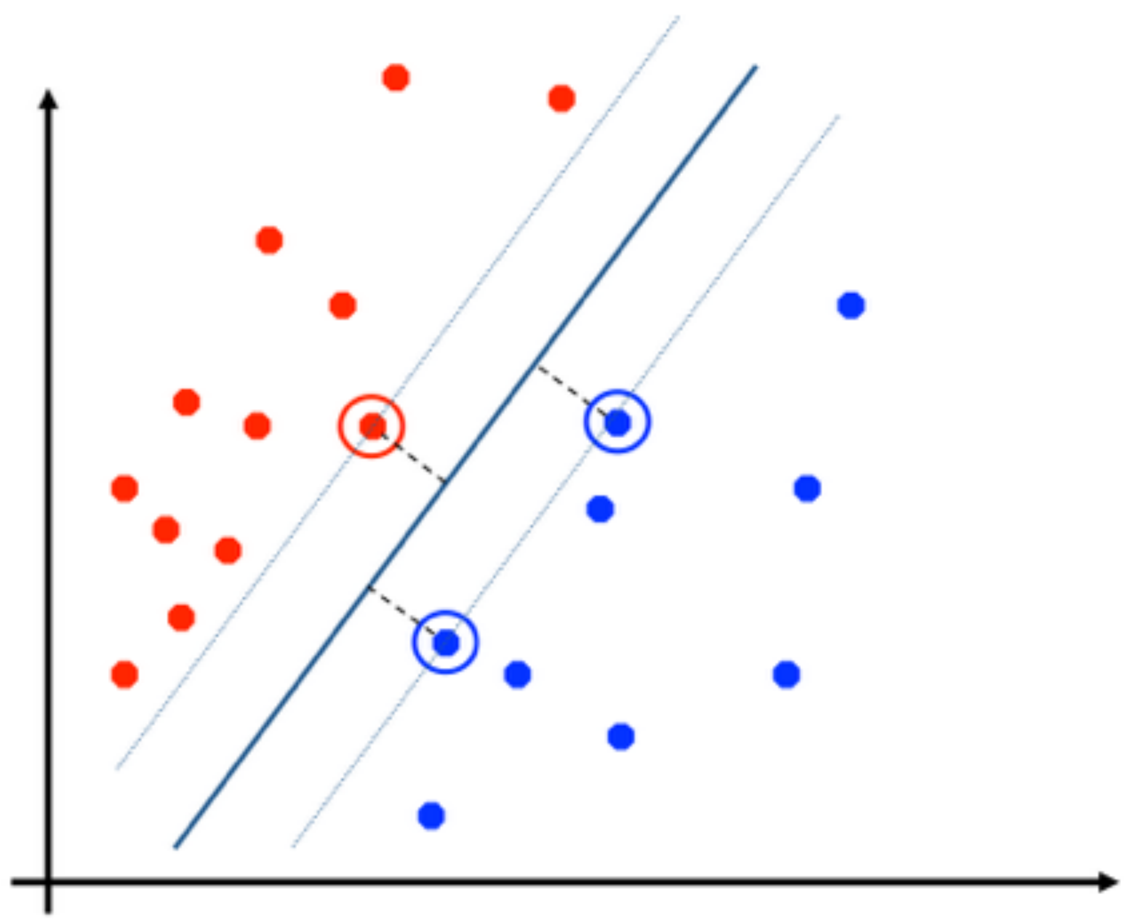
- Each non-zero α_i indicates that corresponding $\mathbf{x}_i$ is a support vector.
- Then the classifying function is (note that we don't need **w** explicitly):

$$\boxed{f(\mathbf{x}) = \Sigma \alpha_i y_i \mathbf{x}_i^{\mathbf{T}} \mathbf{x} + b}$$

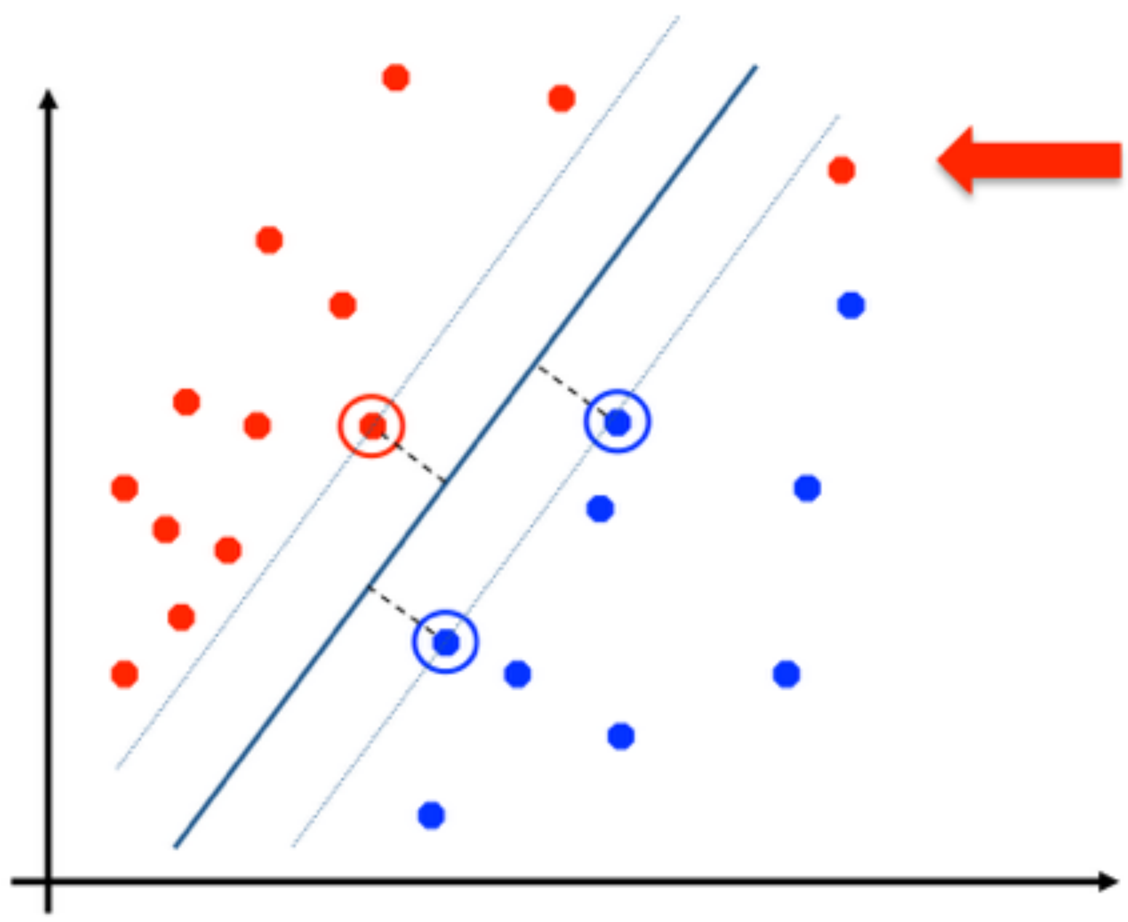
- Notice that it relies on an *inner product* between the test point **x** and the support vectors $\mathbf{x}_i$ – we will return to this later.
- Also keep in mind that solving the optimization problem involved computing the inner products $\mathbf{x}_i^{\mathbf{T}} \mathbf{x}_j$ between all training points.

Questions?

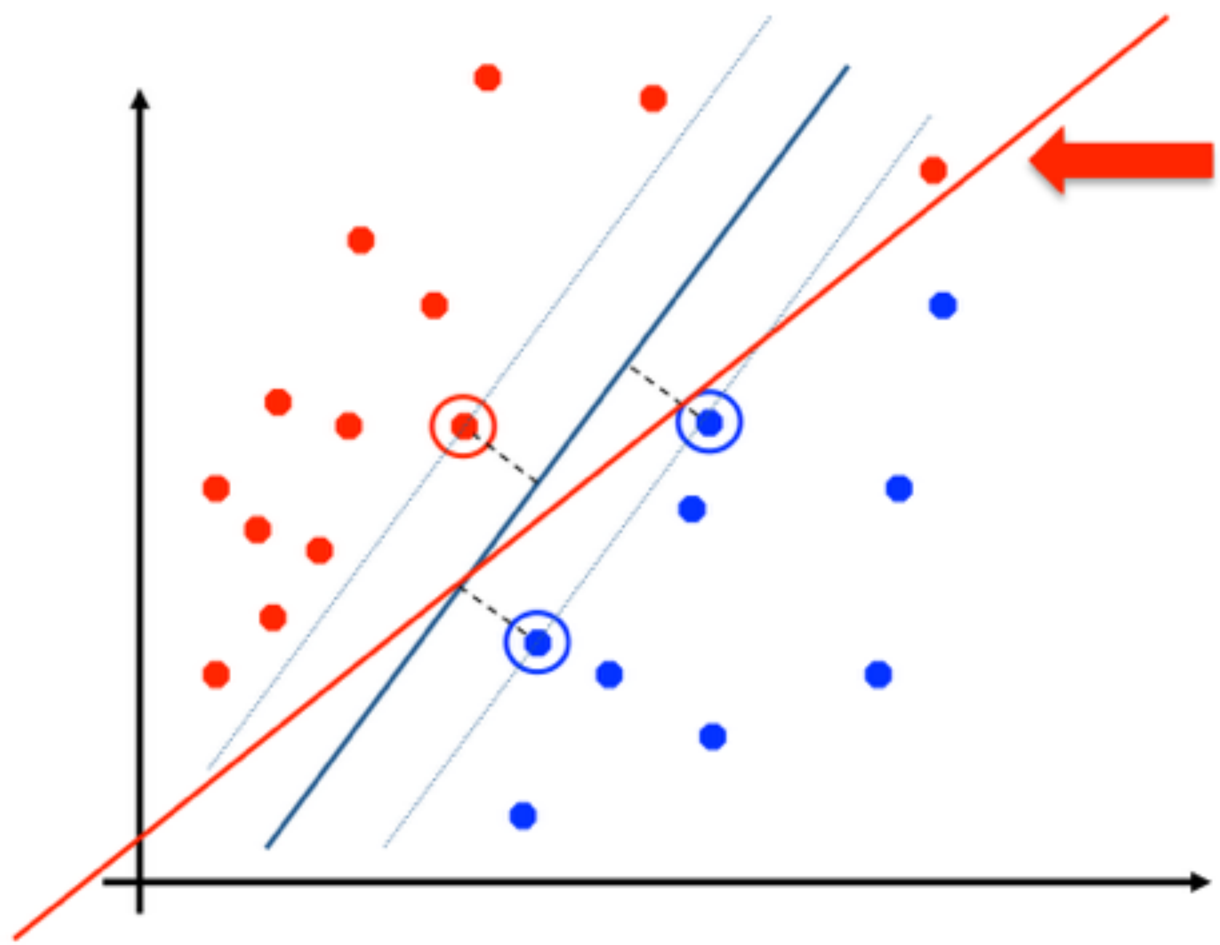
Soft Margin Classification



Soft Margin Classification

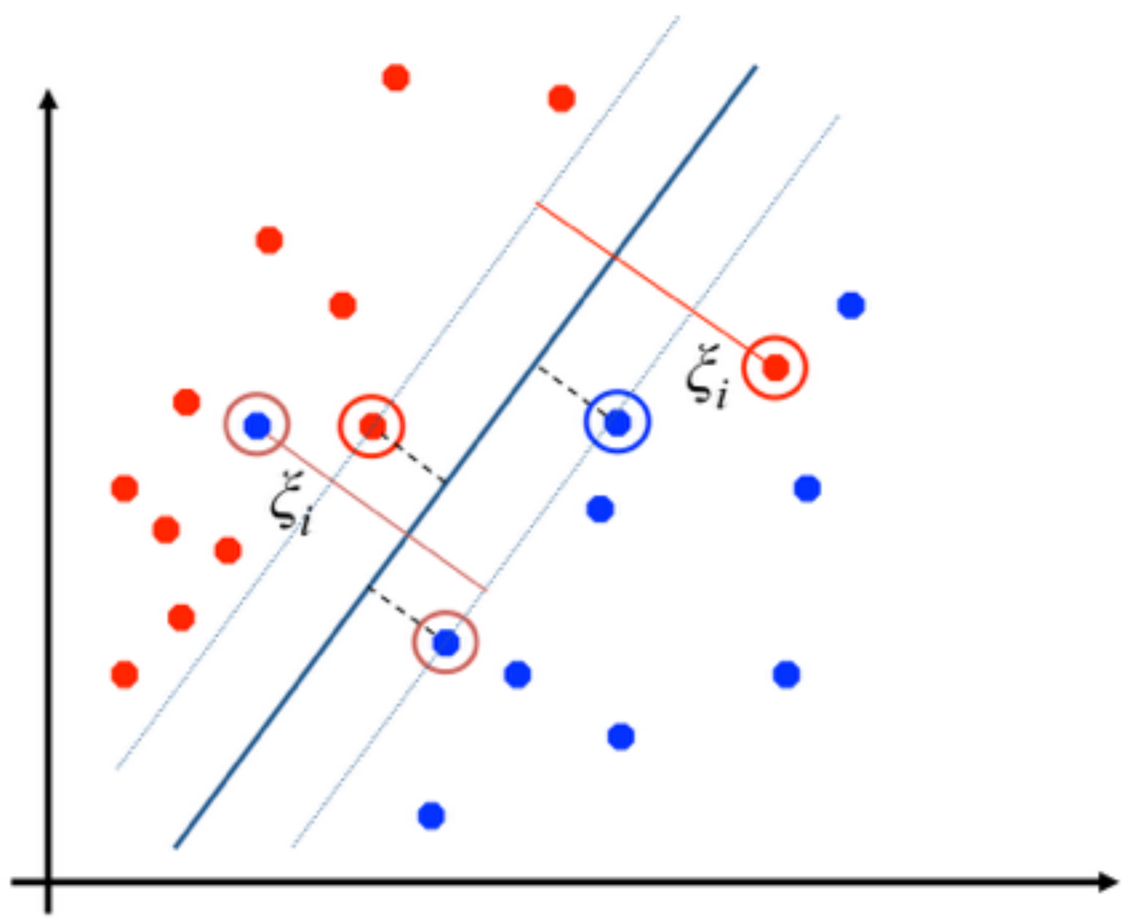


Soft Margin Classification



Soft Margin Classification

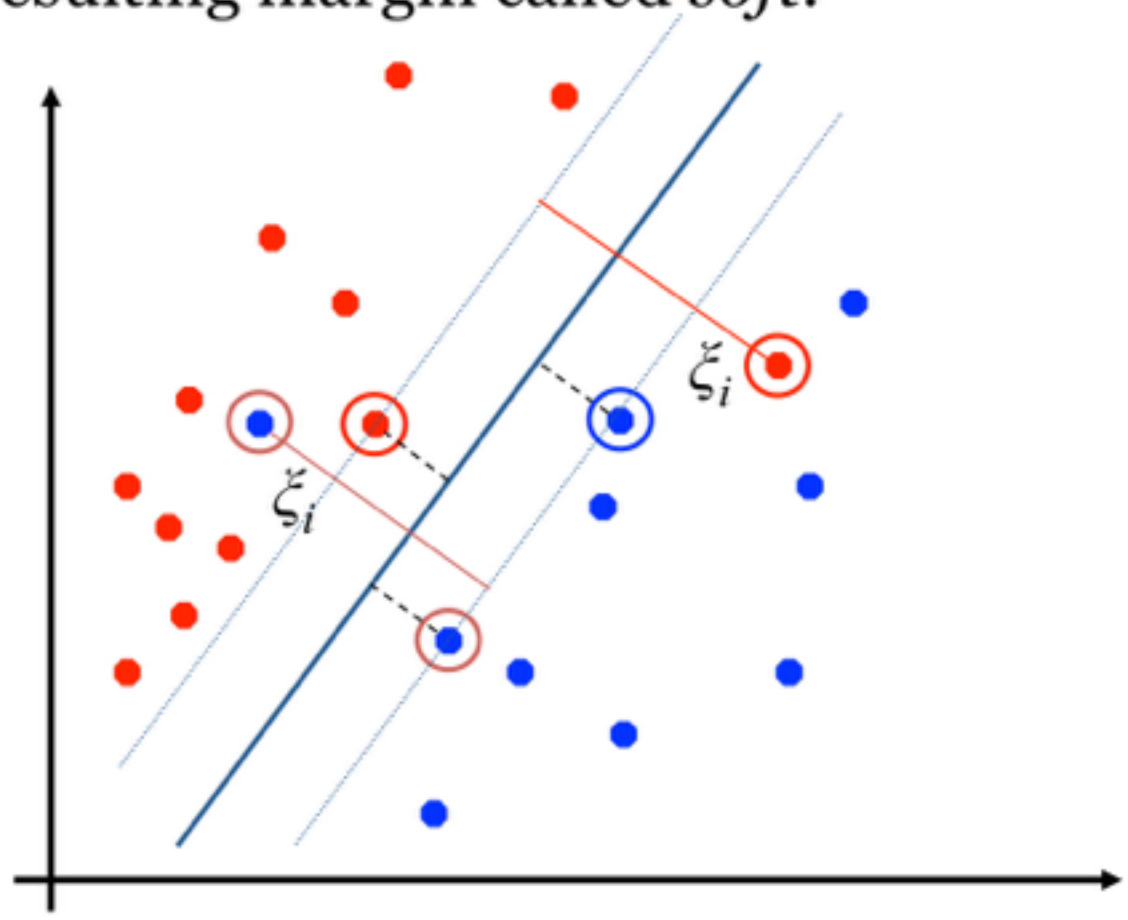
What if the training set is not linearly separable?



Soft Margin Classification

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*.



Soft Margin Classification Mathematically

The old formulation:

Find **w** and b such that
$\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{w}$ is minimized
and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1$

Modified formulation incorporates slack variables:

Find **w** and b such that
$\mathbf{\Phi}(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum\xi_i$ is minimized
and for all $(\mathbf{x}_i, y_i)$, $i=1..n$: $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_{i,}$, $\xi_i \geq 0$

Soft Margin Classification Mathematically

Find $\mathbf{w}$ and b such that

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

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

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.

Small C: less support vectors, maximizing the margin, smooth, underfit
Large C: more support vectors, fitting the train data correctly, overfit

Soft Margin Classification – Solution

The old formulation:

$$\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x}_i$$
$$b = y_k - \Sigma\alpha_i y_i \mathbf{x}_i^{\mathbf{T}} \mathbf{x}_k \quad \text{for any } \alpha_k > 0$$

Modified formulation incorporates slack variables:

$$\mathbf{w} = \Sigma\alpha_i y_i \mathbf{x}_i$$
$$b = y_k(1 - \xi_k) - \Sigma\alpha_i y_i \mathbf{x}_i^{\mathbf{T}} \mathbf{x}_k \quad \text{for any } k \text{ s.t. } \alpha_k > 0$$

Linear SVMs: Summary

The classifier is a *separating hyperplane*.

Most "important" training points are support vectors; they define the hyperplane.

SVMs use quadratic programming via Lagrange multipliers to find the optimal solution for this problem. Training points $\mathbf{x}_i$ with non-zero Lagrangian multipliers α_i are support vectors.