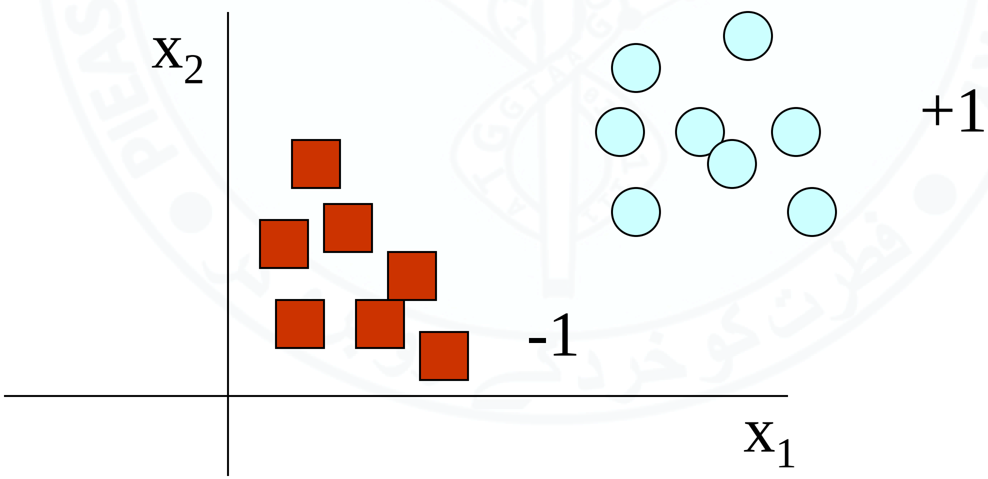
Support Vector Machine Formulation

Dr. Fayyaz ul Amir Afsar Minhas
PIEAS Biomedical Informatics Research Lab
Department of Computer and Information Sciences
Pakistan Institute of Engineering & Applied Sciences
PO Nilore, Islamabad, Pakistan
http://faculty.pieas.edu.pk/fayyaz/

Classification

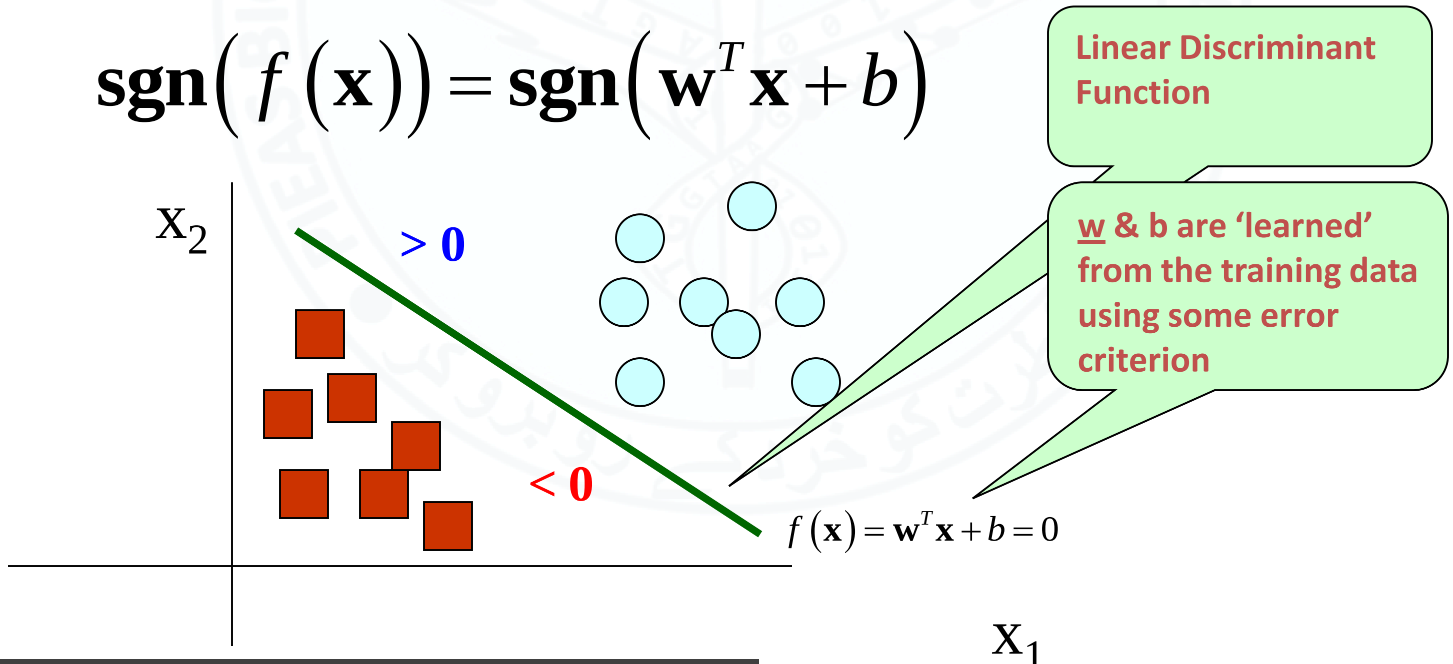
- Before moving on with the discussion let us restrict ourselves to the following problem
  - T = Given Training Set = $\{(\underline{x}^{(i)}, y_i), i = 1 \ldots N\}$
    - $\underline{x}^{(i)} \varepsilon R^m$ {Data Point i }
    - y_i: class of data point i (+1 or -1)



Use of Linear Discriminant in Classification

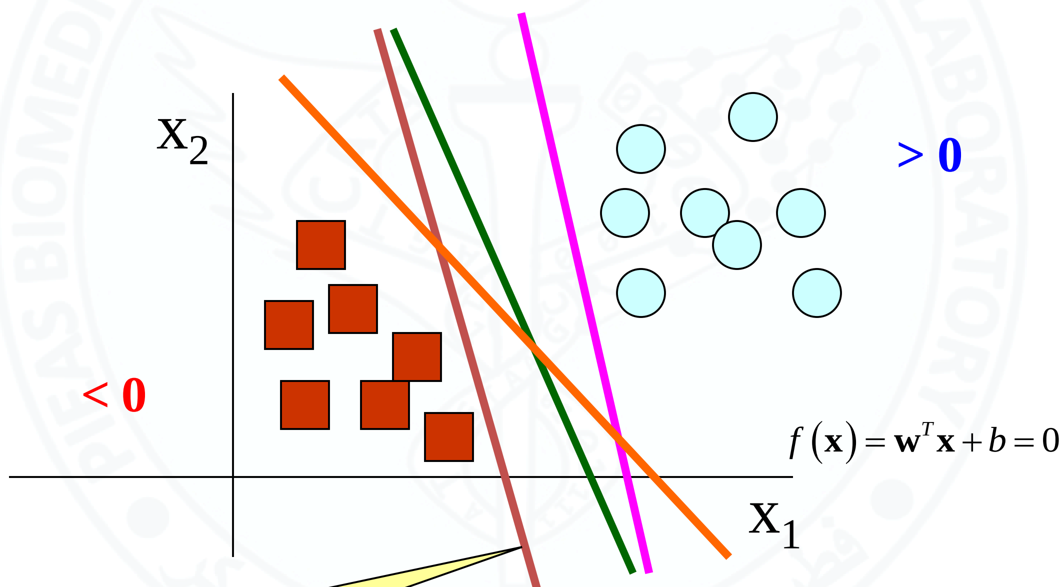
- Classifiers such as the Single Layer Perceptron (with linear activation function) and SVM use a linear discriminant function to differentiate between patterns of different classes
- The linear discriminant function is given by

$$\mathbf{sgn}\left(f(\mathbf{x})\right) = \mathbf{sgn}\left(\mathbf{w}^T\mathbf{x} + b\right)$$



Use of Linear Discriminant in Classification

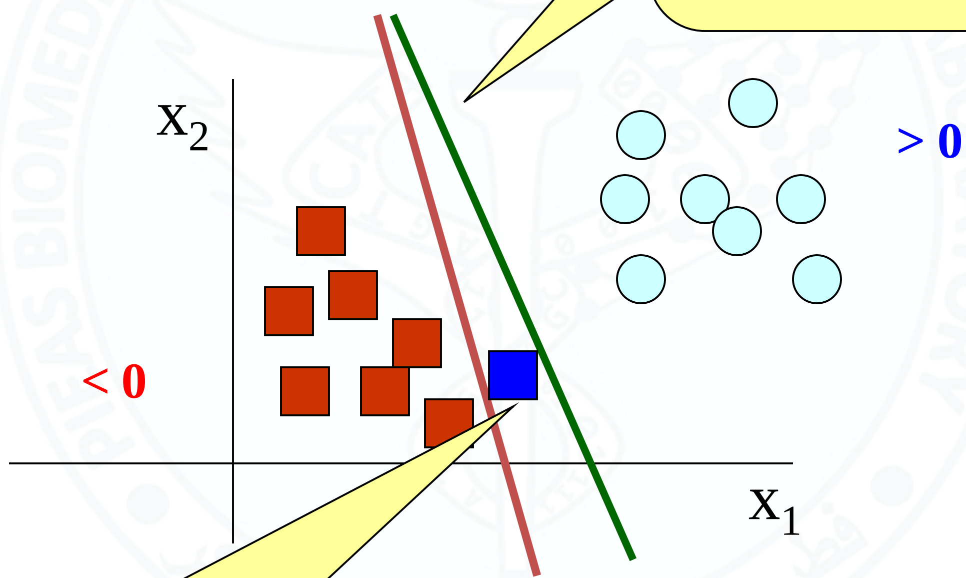
- There are a large number of lines (or in general 'hyperplanes') separating the two classes



Which separator is the best?

Use of Linear Discriminant in Classification

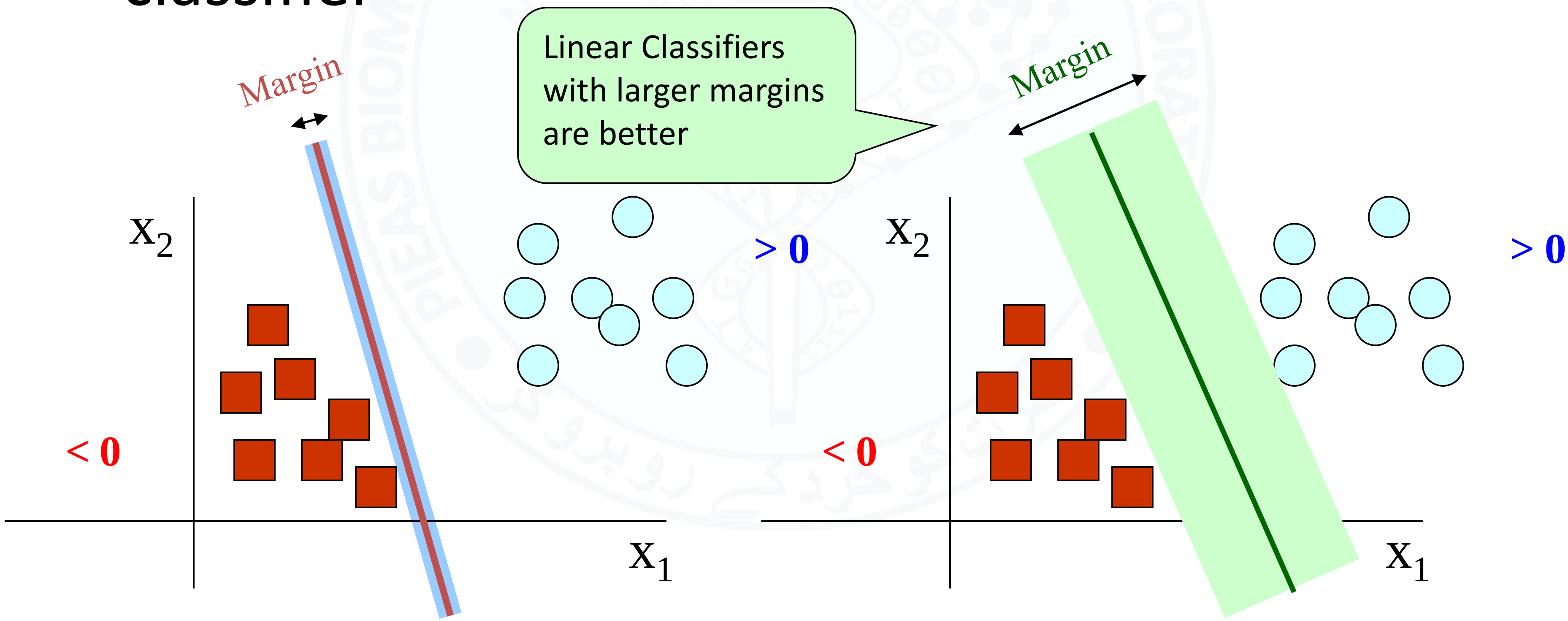
The boundary which lies at the maximum distance from data points of both classes gives better tolerance to noise and better generalization



Misclassified as -1 by the 'red' discriminant

Margin of a linear classifier

- The width by which the boundary of a linear classifier can be increased before hitting a data point is called the margin of the linear classifier



Support Vector Machines (SVM)

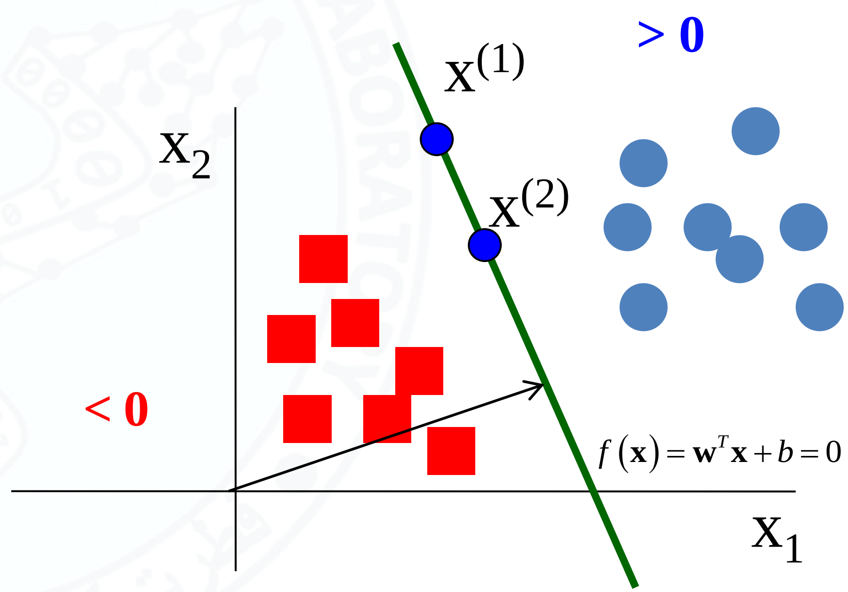
- Support Vector Machines are linear classifiers that produce the optimal separating boundary (hyper-plane)
  - Find **w** and b in a way so as to maximize the margin while classifying all the training patterns correctly (for linearly separable problem)

Finding Margin of a Linear Classifier

- Consider a linear classifier with the boundary

$$f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0 \quad \textbf{for all x on the boundary}$$

- We know that the vector w is perpendicular to the boundary
  - Consider two points $x^{(1)}$ and $x^{(2)}$ on the boundary



$$f\left(\mathbf{x}^{(1)}\right) = \mathbf{w}^T\mathbf{x}^{(1)} + b = 0 \qquad (1)$$

$$f\left(\mathbf{x}^{(2)}\right) = \mathbf{w}^T\mathbf{x}^{(2)} + b = 0 \qquad (2)$$

Subtracting (1) from (2)

$$\mathbf{w}^T\left(\mathbf{x}^{(2)} - \mathbf{x}^{(1)}\right) = 0 \quad \Rightarrow \quad \mathbf{w} \perp \left(\mathbf{x}^{(2)} - \mathbf{x}^{(1)}\right)$$

Finding Margin of a Linear Classifier

- Let $x^{(s)}$ be a point in the feature space with its projection $x^{(p)}$ on the boundary

$$f\left(\mathbf{x}^{(p)}\right) = \mathbf{w}^T\mathbf{x}^{(p)} + b = 0$$

- We know that,

$$\mathbf{x}^{(s)} = \mathbf{x}^{(p)} + r\hat{\mathbf{w}}$$

$$\Rightarrow f\left(\mathbf{x}^{(s)}\right) = \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(s)} + b$$

$$= \mathbf{w}^{\mathrm{T}}\left(\mathbf{x}^{(p)} + r\hat{\mathbf{w}}\right) + b$$

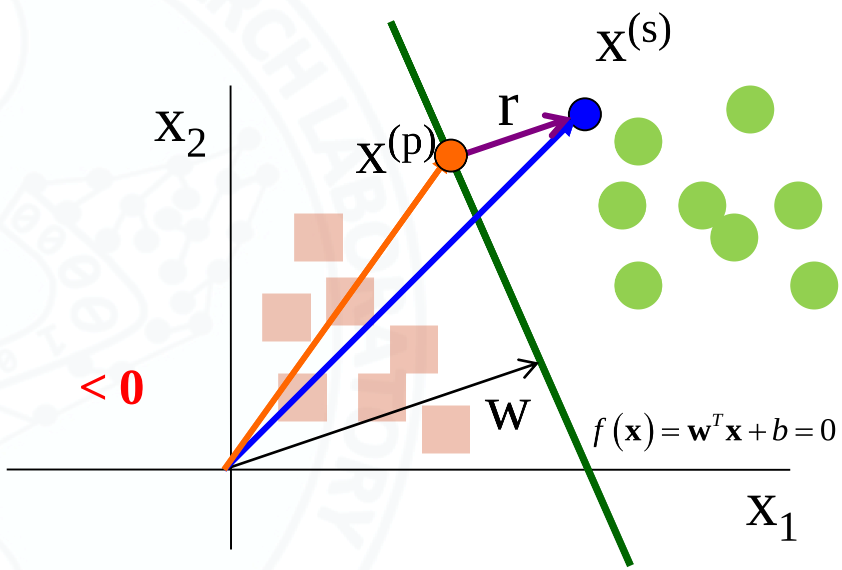
$$= \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(p)} + r\mathbf{w}^{\mathrm{T}}\hat{\mathbf{w}} + b$$

$$= \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(p)} + b + r\mathbf{w}^{\mathrm{T}}\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

$$= 0 + r\|\mathbf{w}\| = r\|\mathbf{w}\|$$

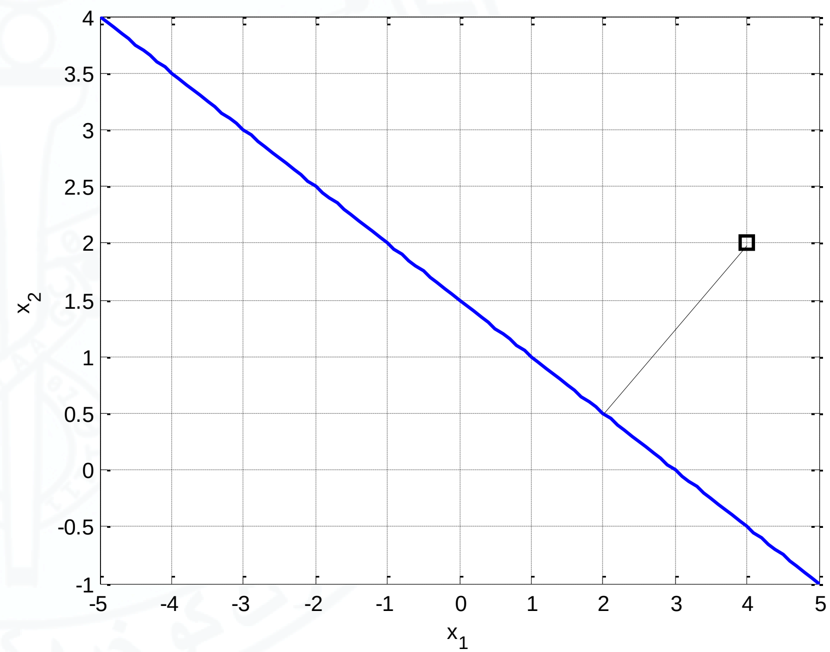
$$\Rightarrow r = \frac{f\left(\mathbf{x}^{(s)}\right)}{\|\mathbf{w}\|}$$

Perpendicular Distance of a point from a line



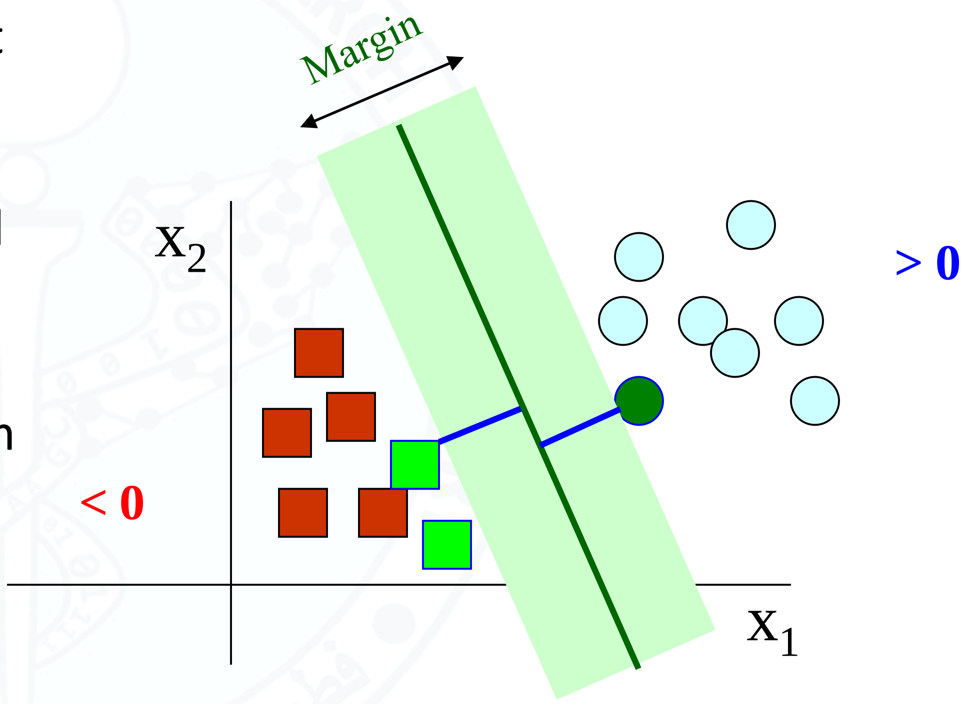
Example

- Consider the line
  - $x_1+2x_2+3 = 0$
- The distance of (4,2) is
  - r = 4.92



Finding Margin of a Linear Classifier

- The points in the training set that lie closest (having minimum perpendicular distance) to the separating hyper-plane are called support vectors
- Margin is twice of perpendicular distances of the hyper-plane from the nearest support vector of the two classes

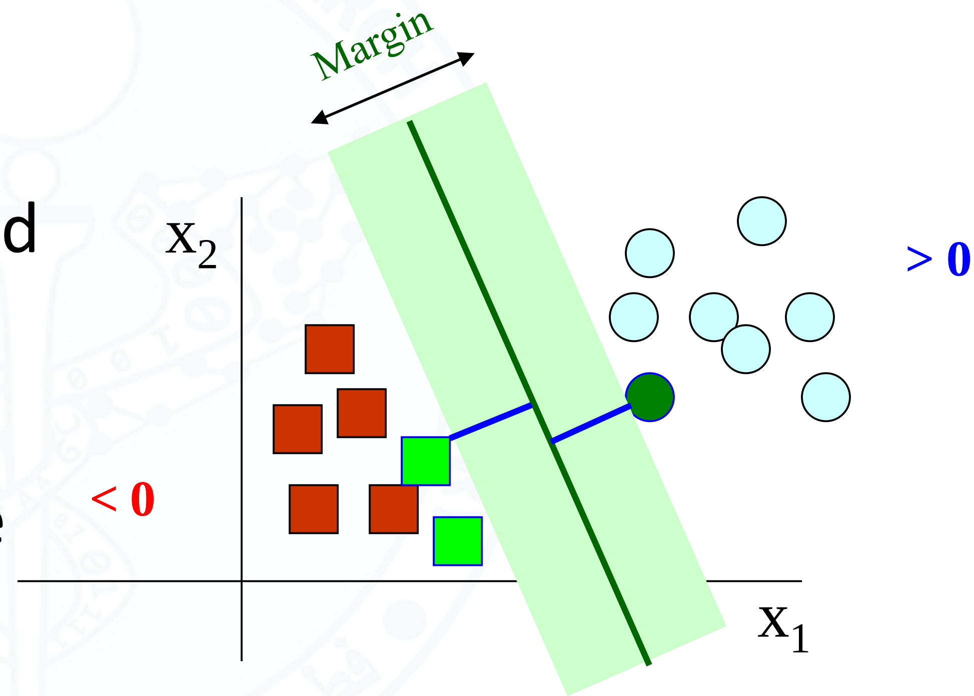


$$\rho = 2|r| = 2\left|\frac{f\left(\mathbf{x}^*\right)}{\|\mathbf{w}\|}\right|$$

Nearest Support Vector

Geometric vs. Functional Margin

- Functional Margin
  - This gives the position of the point with respect to the plane, which does not depend on the magnitude.
- Geometric Margin
  - This gives the distance between the given training example and the given plane.



Finding Margin of a Linear Classifier

- Assume that all data is at least distance 1 from the hyperplane, then the following two constraints follow for a training set $\{(x^{(i)}, y_i)\}$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b \geq 1 \quad \text{if } y_i = 1$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b \leq -1 \quad \text{if } y_i = -1$$

$$\Rightarrow \rho = 2|r| = 2\left|\frac{f\left(\mathbf{x}^*\right)}{\|\mathbf{w}\|}\right| = 2\left|\frac{\pm 1}{\|\mathbf{w}\|}\right|$$

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

Support Vector Machines

- Support Vector Machines, in their basic form, are linear classifiers that are trained in a way so as to maximize the margin
- Principles of Operation
  - Define what an optimal hyper-plane is (in way that can be identified in a computationally efficient way)
    - Maximize margin
      - Allows noise tolerance
  - Extend the above definition for non-linearly separable problems
    - have a penalty term for misclassifications
  - Map data to an alternate space where it is easier to classify with linear decision surfaces
    - reformulate problem so that data is mapped implicitly to this space (using kernels)

Margin Maximization in SVM

- We know that if we require

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b \geq 1 \quad \text{if } y_i = 1$$

$$\mathbf{w}^{\mathbf{T}}\mathbf{x}^{(\mathbf{i})} + b \leq -1 \quad \text{if } y_i = -1$$

- Then the margin is

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

- Margin maximization can be performed by reducing the norm of the w vector

SVM as an Optimization problem

- We can present SVM as the following optimization problem

$$\textbf{max} \quad \rho = \frac{2}{\|\mathbf{w}\|}$$

$$s.t.$$

$$\mathbf{w}^T \mathbf{x}^{(i)} + b \leq -1, \qquad \forall i \ \textbf{s.t.}\ y_i = -1$$

$$\mathbf{w}^T \mathbf{x}^{(i)} + b \geq +1, \qquad \forall i \ \textbf{s.t.}\ y_i = +1$$

OR

$$\textbf{min} \quad \rho = \frac{1}{2}\mathbf{w}^T \mathbf{w}$$

$$s.t.$$

$$y_i \left(\mathbf{w}^T \mathbf{x}^{(i)} + b \right) \geq +1$$