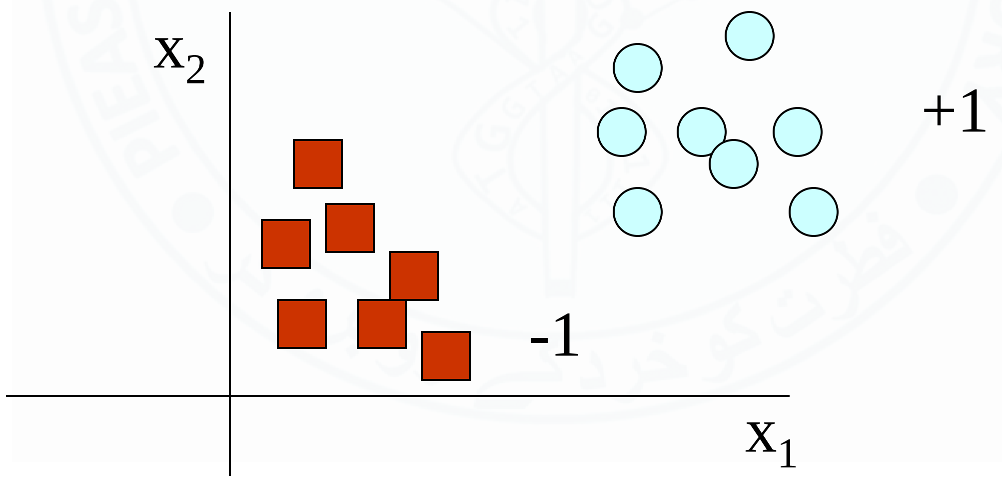
Support Vector Machines

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)



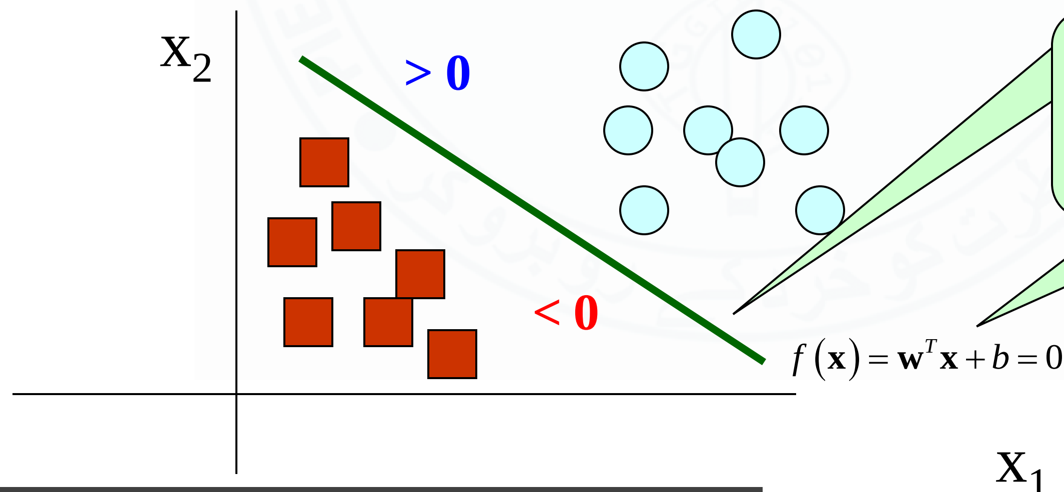
Use of Linear Discriminant in Classification

- 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

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

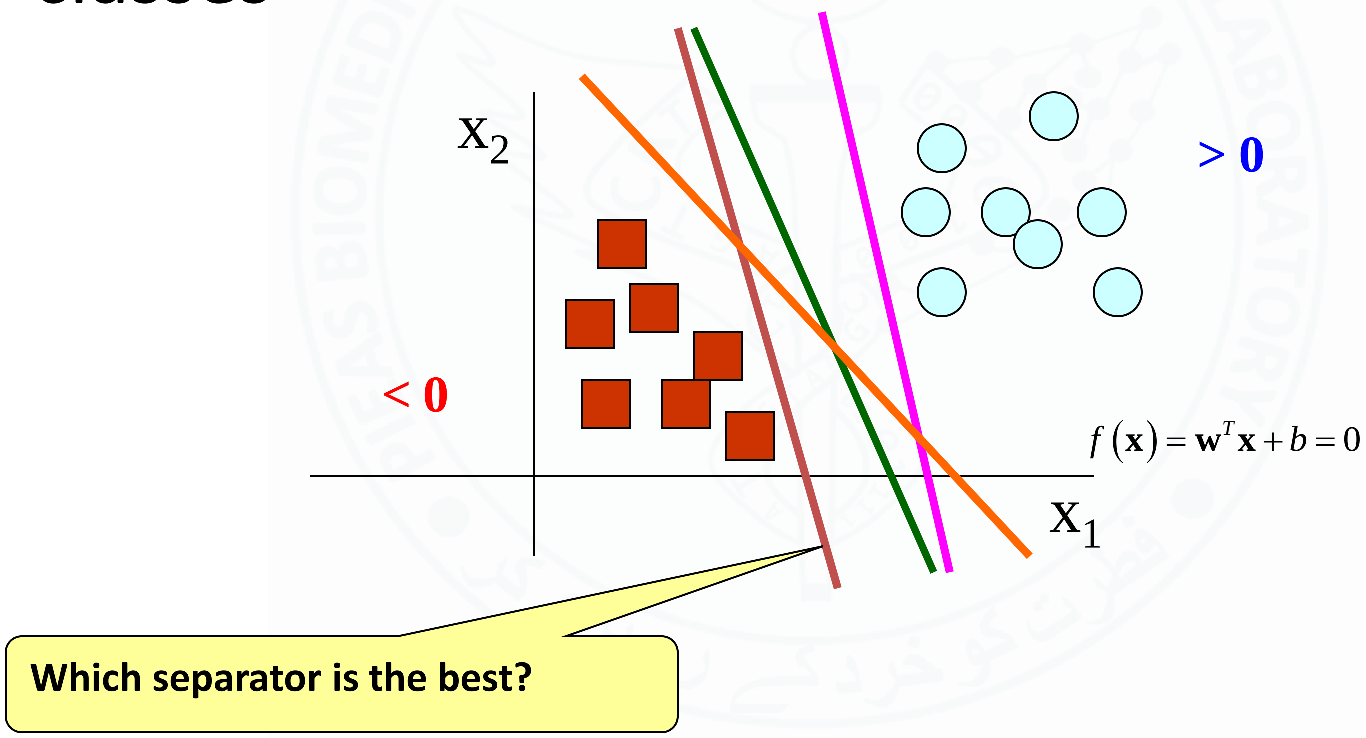
Linear Discriminant Function

<u>w</u> & b are ‘learned’ from the training data using some error criterion

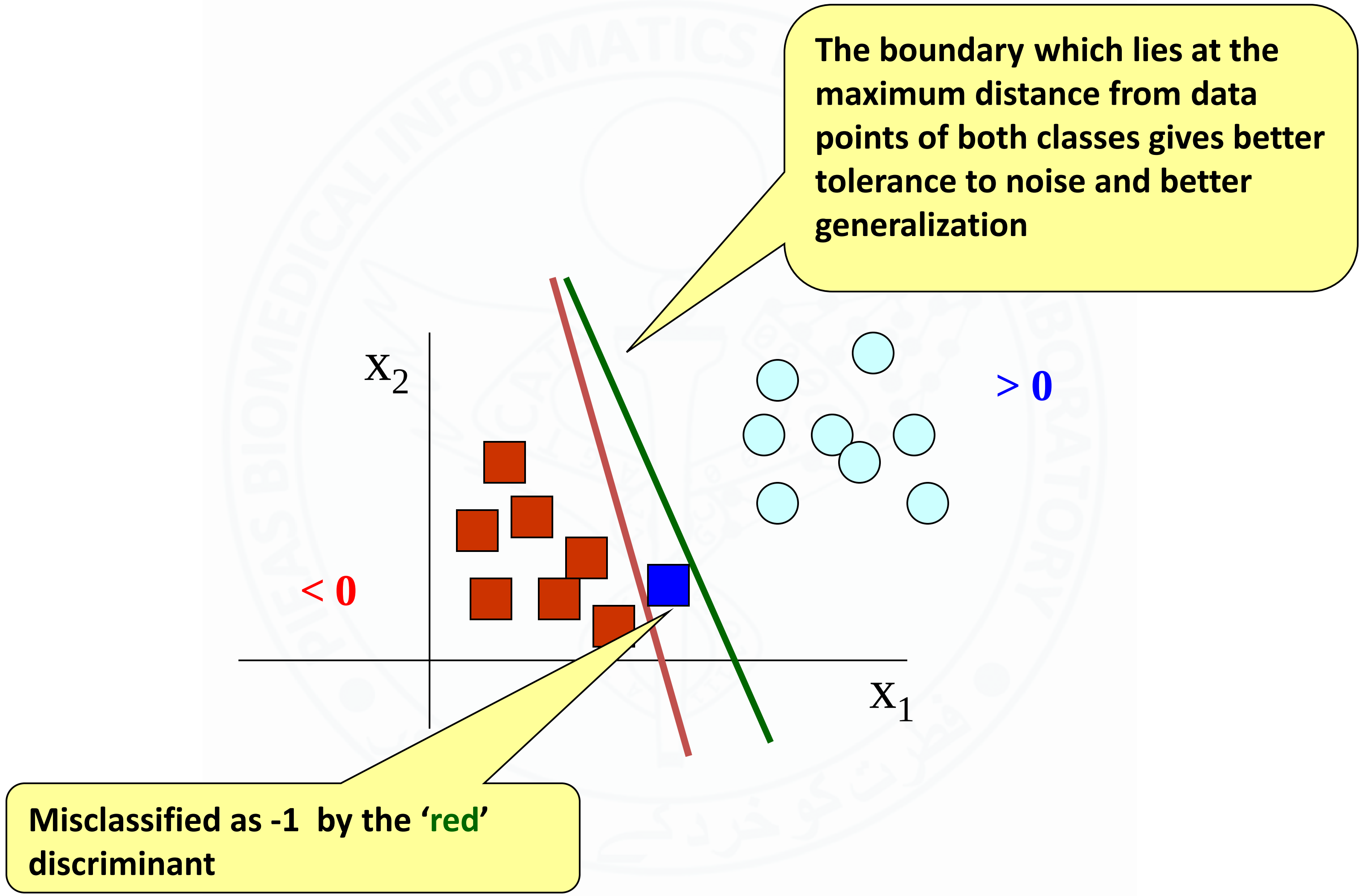


Use of Linear Discriminant in Classification

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

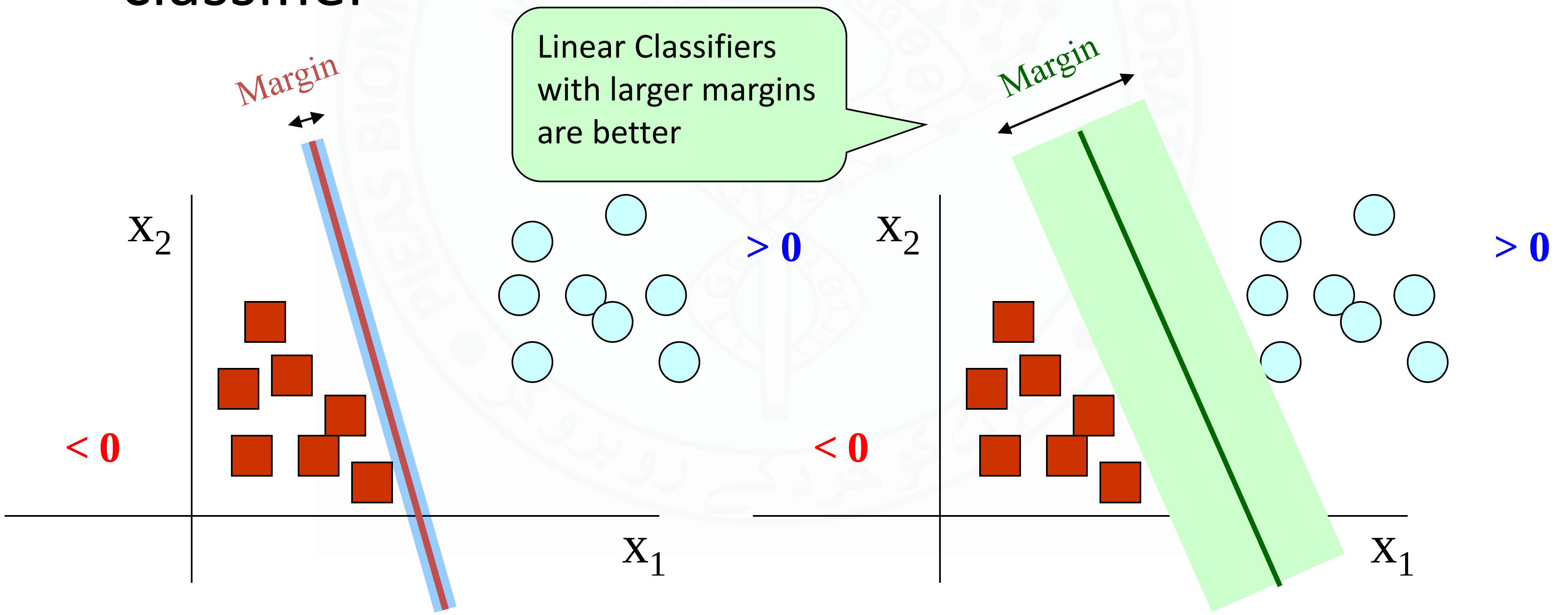


Use of Linear Discriminant in Classification



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)

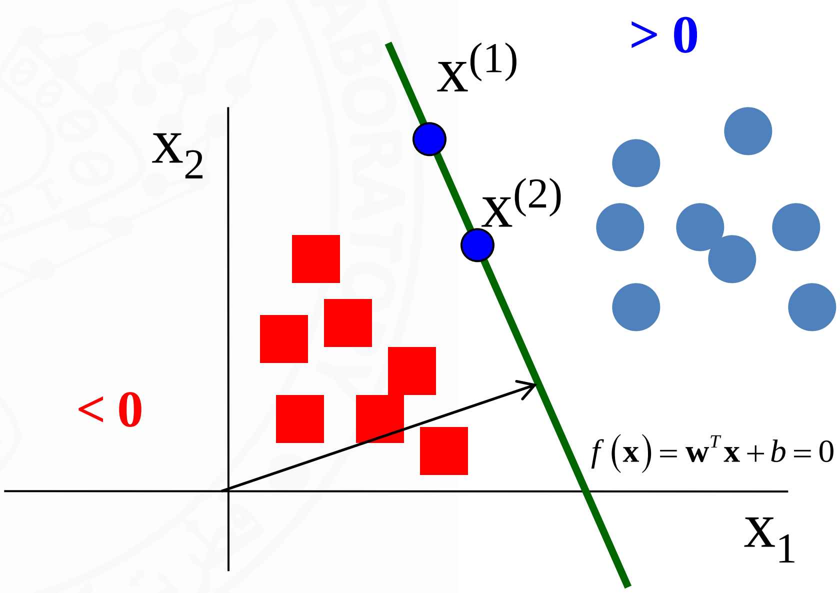
- Support Vector Machines are linear classifiers that produce the optimal separating boundary (hyper-plane)
  - Find **w** and b in a way so as to maximize the margin while classifying all the training patterns correctly (for linearly separable problem)

Finding Margin of a Linear Classifier

- Consider a linear classifier with the boundary

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

- 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

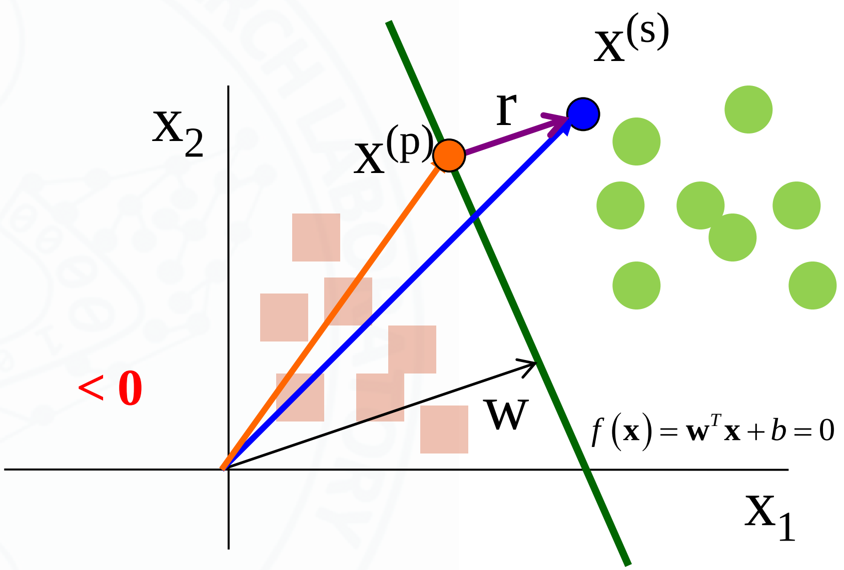
$$f\left(\mathbf{x}^{(p)}\right) = \mathbf{w}^T\mathbf{x}^{(p)} + b = 0$$

- We know that,

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

$$\begin{aligned}
\Rightarrow f\left(\mathbf{x}^{(s)}\right) &= \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(s)} + b \\
&= \mathbf{w}^{\mathrm{T}}\left(\mathbf{x}^{(p)} + r\hat{\mathbf{w}}\right) + b \\
&= \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(p)} + r\mathbf{w}^{\mathrm{T}}\hat{\mathbf{w}} + b \\
&= \mathbf{w}^{\mathrm{T}}\mathbf{x}^{(p)} + b + r\mathbf{w}^{\mathrm{T}}\frac{\mathbf{w}}{\|\mathbf{w}\|} \\
&= 0 + r\|\mathbf{w}\| = r\|\mathbf{w}\|
\end{aligned}$$

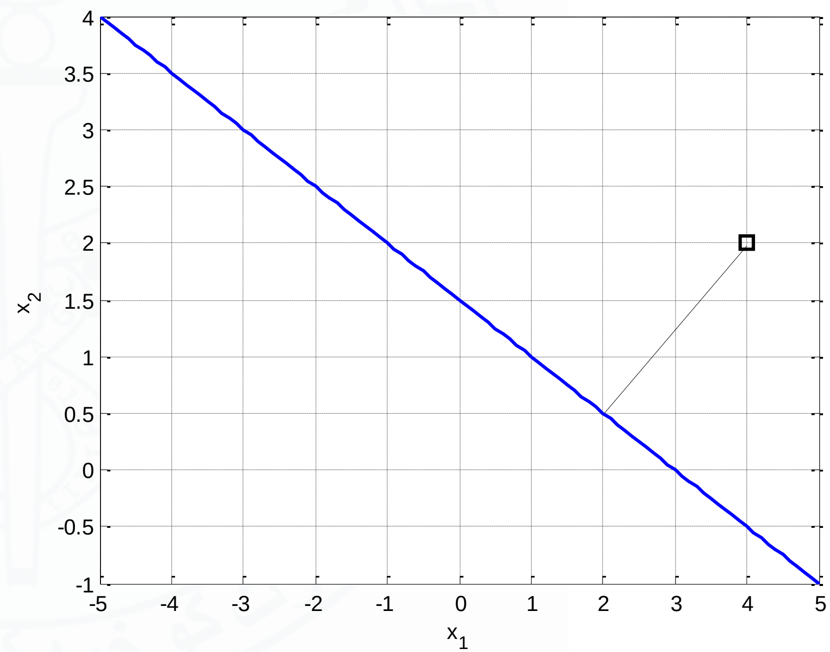
$$\Rightarrow r = \frac{f\left(\mathbf{x}^{(s)}\right)}{\|\mathbf{w}\|}$$



Perpendicular Distance of a point from a line

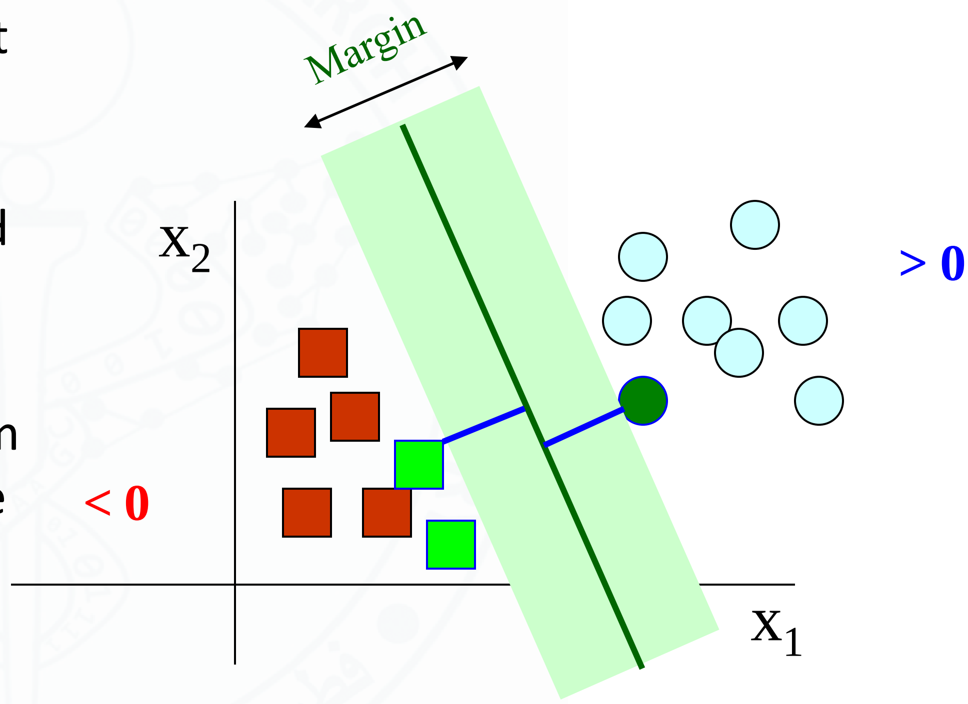
Example

- 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

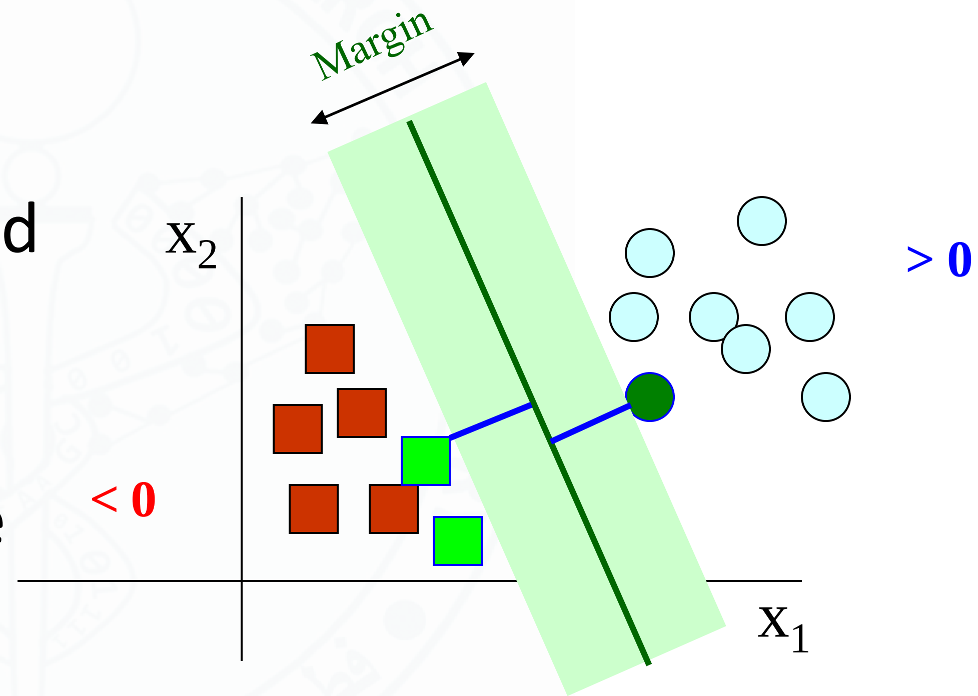


$$\rho = 2|r| = 2\left|\frac{f\left(\mathbf{x}^{*}\right)}{\|\mathbf{w}\|}\right|$$

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

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

$$s.t.$$

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

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

OR

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

$$s.t.$$

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

Example: Solution of the OR problem

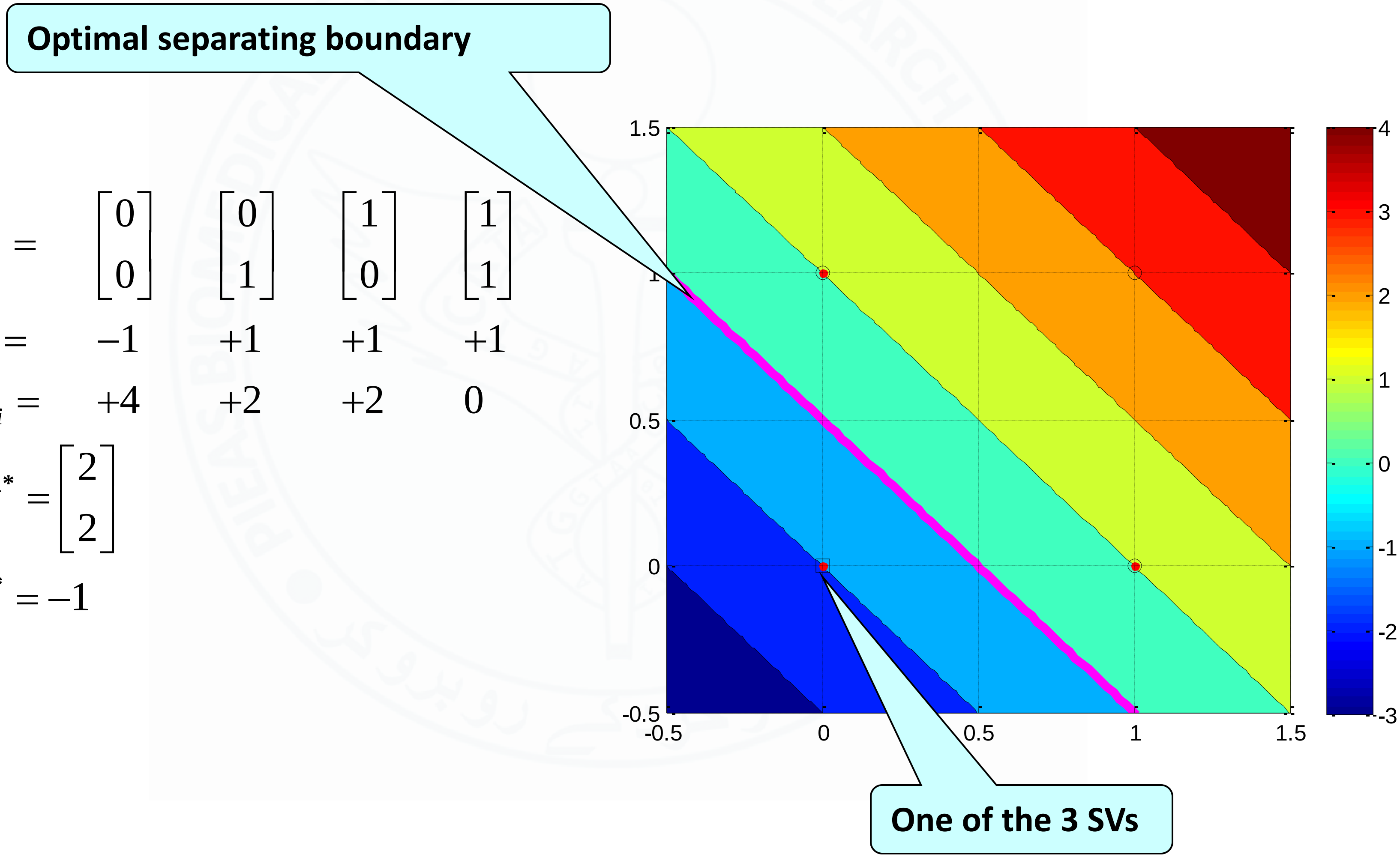
$$\mathbf{x}^i = \quad \begin{bmatrix}0\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\end{bmatrix} \quad \begin{bmatrix}1\\0\end{bmatrix} \quad \begin{bmatrix}1\\1\end{bmatrix}$$

$$y = \quad -1 \quad +1 \quad +1 \quad +1$$

$$\alpha_i = \quad +4 \quad +2 \quad +2 \quad 0$$

$$\mathbf{w}^* = \begin{bmatrix}2\\2\end{bmatrix}$$

$$b^* = -1$$



Handling Non-Separable Patterns

- All the derivation till now was based on the requirement that the data be linearly separable
  - Practical Problems are Non-Separable
- Non-separable data can be handled by relaxing the constraint

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

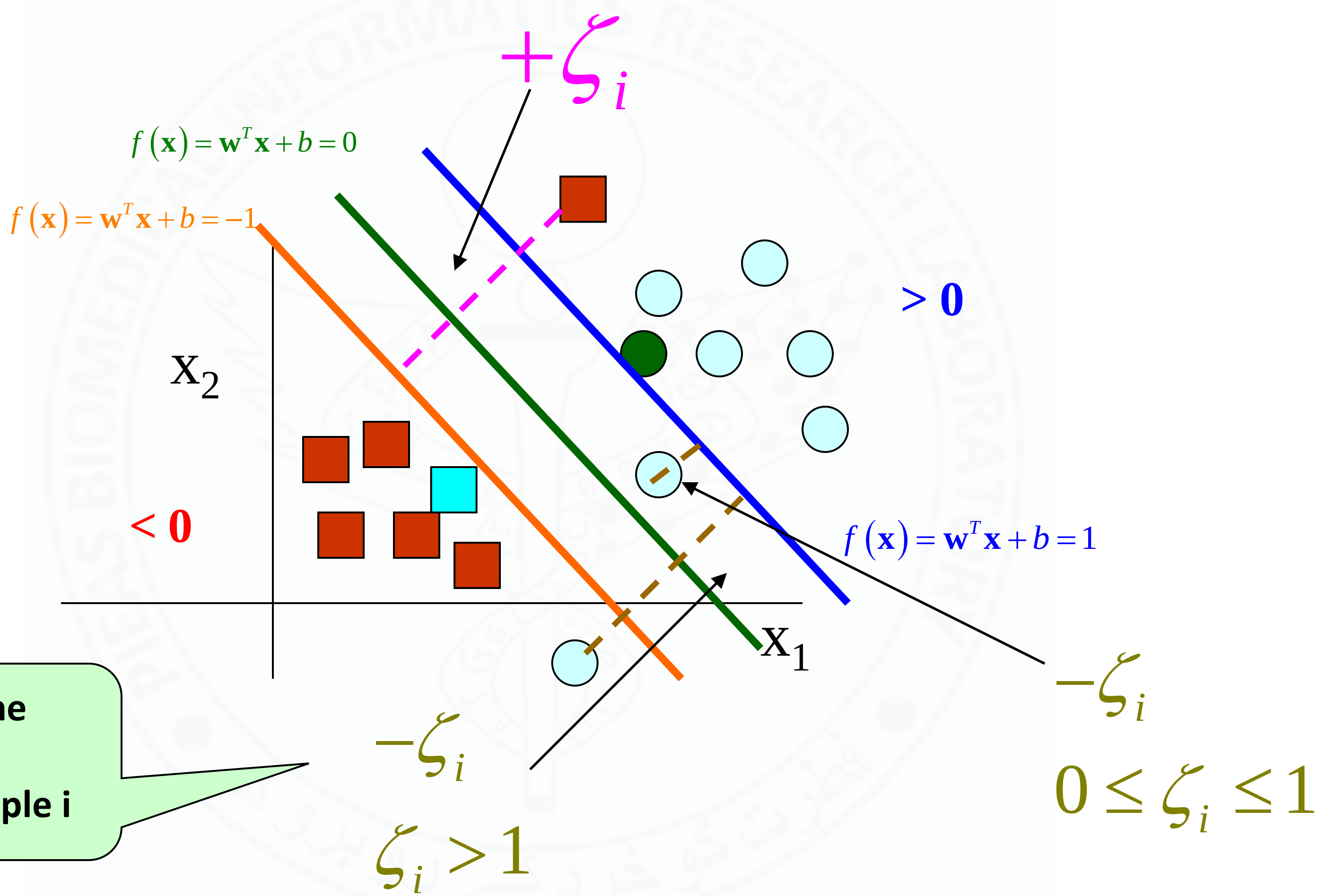
- This is achieved as

$$\left.\begin{array}{ll}\mathbf{w}^T\mathbf{x}^{(i)}+b\geq 1-\zeta_i & \forall y_i=+1\\ \mathbf{w}^T\mathbf{x}^{(i)}+b\geq -1+\zeta_i & \forall y_i=-1\end{array}\right\} \quad y_i\left(\mathbf{w}^T\mathbf{x}^{(i)}+b\right)\geq 1-\zeta_i$$

$$\zeta_i\geq 0$$

Slack Variables

Handling Non-Separable Patterns (Soft Margin)



ζ_i tells us about the extent of margin violation of example i

Handling Non-Separable Patterns (Soft Margin)

- Our objective in designing a SVM here is to maximize the margin while minimizing the misclassification error
- The misclassification error can be written as:

$$\Phi(\zeta) = \sum_{i=1}^{N} I(\zeta_i - 1) \qquad I(\zeta) = \begin{cases} 0 & \zeta \le 0 \\ 1 & else \end{cases}$$

- Since this function is non-linear and non-convex, therefore we choose to use an approximate given by

$$\Phi(\zeta) = \sum_{i=1}^{N} \zeta_i$$

Soft Margin SVM as an Optimization Problem

- The overall optimization problem can be written as

$$
\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\zeta_i \\
s.t. \quad & y_i\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \geq 1 - \zeta_i \\
& \zeta_i \geq 0
\end{aligned}
$$

Weight of the penalty due to margin violation

Or

$$
\begin{aligned}
\min_{\mathbf{w},b} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{N}\zeta_i \\
s.t. \quad & 1 - \zeta_i - y_i\left(\mathbf{w}^T\mathbf{x}^{(i)} + b\right) \leq 0 \\
& -\zeta_i \leq 0
\end{aligned}
$$

The objective function now optimizes two conflicting objectives: Maximization of the margin and minimization of the margin violations

Example

$$\mathbf{x}^i = \begin{bmatrix}0\\0\end{bmatrix} \quad \begin{bmatrix}0\\1\end{bmatrix} \quad \begin{bmatrix}1\\0\end{bmatrix} \quad \begin{bmatrix}1\\1\end{bmatrix} \quad \begin{bmatrix}0.8\\0.5\end{bmatrix}$$

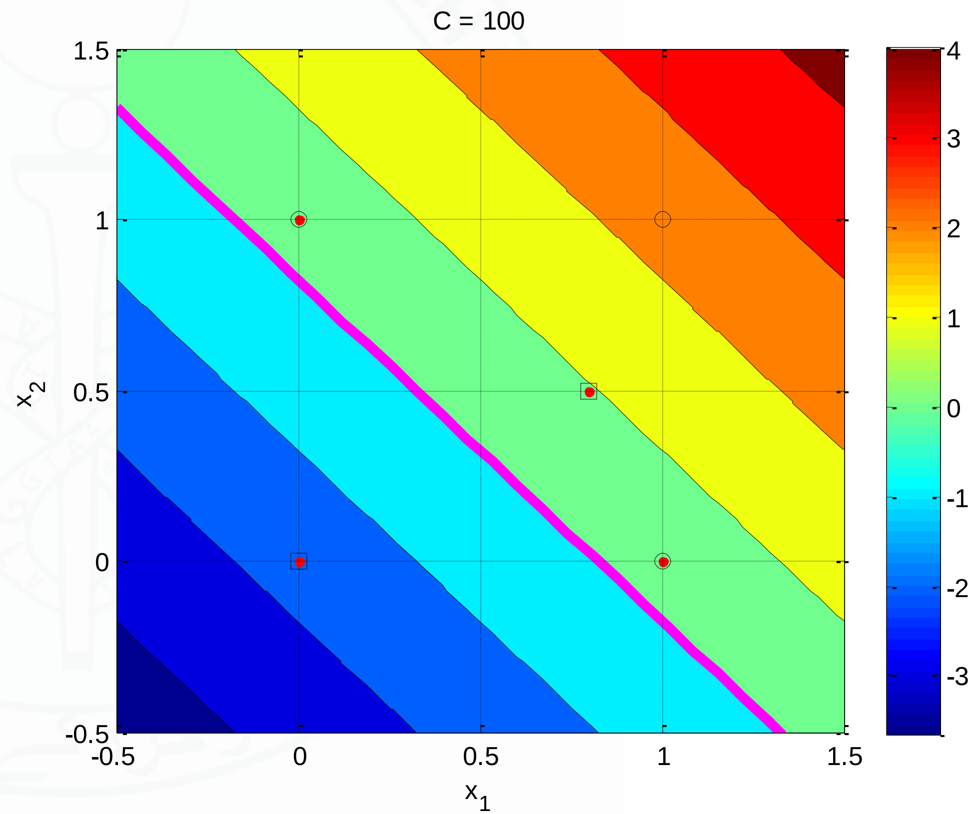
$$C = 100$$

$$y = \quad -1 \quad +1 \quad +1 \quad +1 \quad -1$$

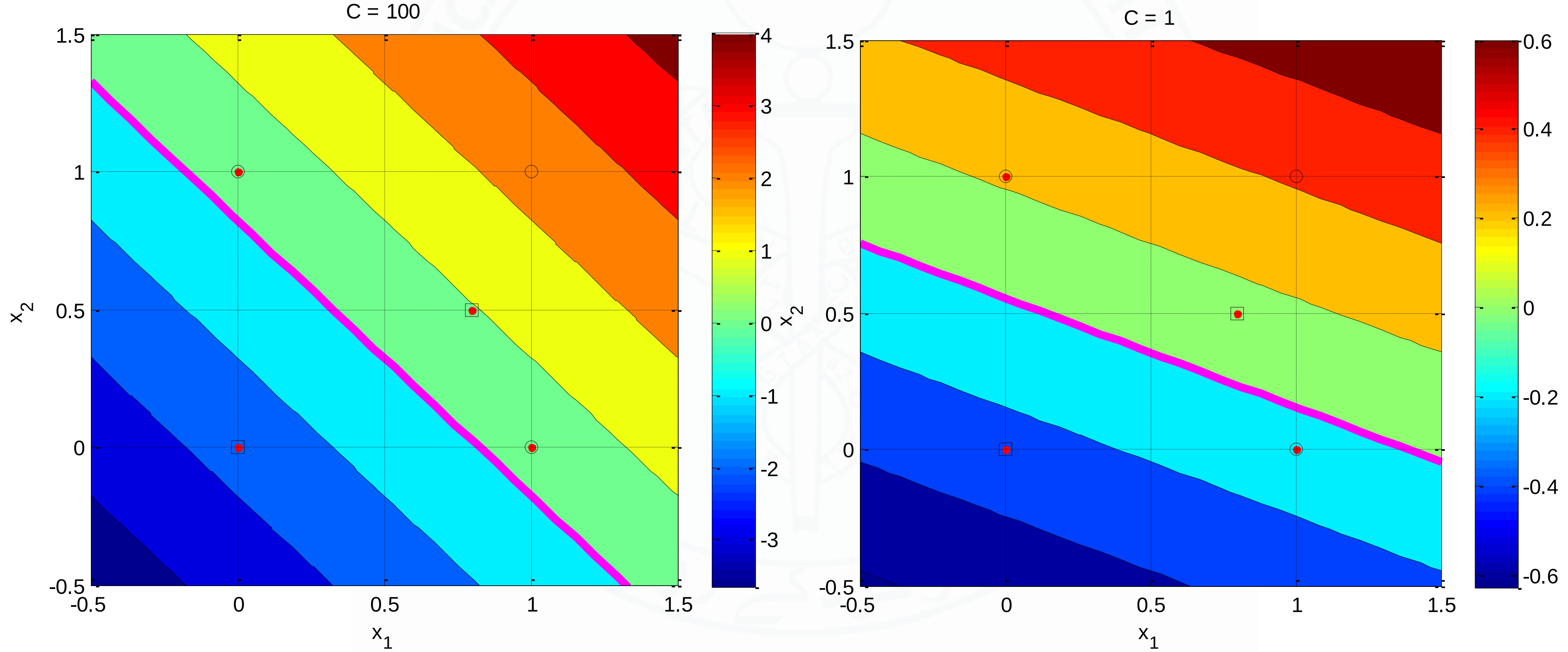
$$\alpha_i = \quad +34 \quad +52 \quad +82 \quad 0 \quad +100$$

$$\mathbf{w}^* = \begin{bmatrix}2\\2\end{bmatrix}$$

$$b^* = -1.65$$

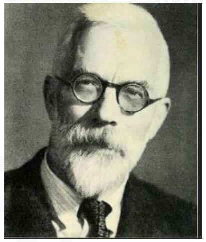


Example…



SVMs uptil now

- Vapnik and Chervonenkis:
  - Hard SVM (1962)
  - Theoretical foundations for SVMs
  - Structural Risk Minimization
- Corinna Cortes
  - Soft SVM (1995)
- Enter: Bernard Scholkopf (1997)
  - Representer Theorem
  - Complete Kernel trick!
  - Kernels not only allow nonlinear boundaries but also allow representation of non-vectoral data


R. A. Fisher
1890-1962


Rosenblatt
1928-1971


V. Vapnik
1936 -


Chervonenkis
1938 - 2014


C. Cortes
1961 -


Scholkopf
1968 -

http://www.svms.org/history.html

Reading

- Sections 10.1-10.3
- Sections 13.1-13.3
- Alpaydin, Ethem. Introduction to Machine Learning. Cambridge, Mass. MIT Press, 2010.