Lecture 2
Linear Classifiers

Professor Slobodan Ribarić
University of Zagreb, Faculty of Electrical Engineering and Computing (FER)
Croatia

GIAN course, March 2018

Linear decision functions

A general linear decision function:

$d(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_nx_n + w_{n+1}$ – linear combinations of feature (pattern) vector's components

w_i ; $i = 1, 2, \ldots, n$ – weights

w_{n+1} – bias, threshold weight

$\mathbf{w}_0 = [w_1, w_2, \ldots , w_n]^T$ – weight vector

$\mathbf{x} = [x_1, x_2, \ldots , x_n]^T$

$$d(\mathbf{x}) = \mathbf{w}_0^T \mathbf{x} + w_{n+1}$$

$\mathbf{x} = [x_1, x_2, \ldots , x_n , 1]^T$ and $\mathbf{w} = [w_1, w_2, \ldots , w_n , w_{n+1}]^T$ – augmented vectors

$$\boxed{d(\mathbf{x}) = \mathbf{w}^T \mathbf{x}}$$

Linear decision function for feature vector $\mathbf{x} = [x_{1,} x_{2,} \dots , x_n]^T$ is:

- for $n = 2$ ….. line
- for $n = 3$ …..plane
- for $n > 3$ ….. hyperplane

Linear decision functions for M = 2 classes (ω_1 and ω_2)

Boundary between two subspaces which correspond to classes ω_1 and ω_2:

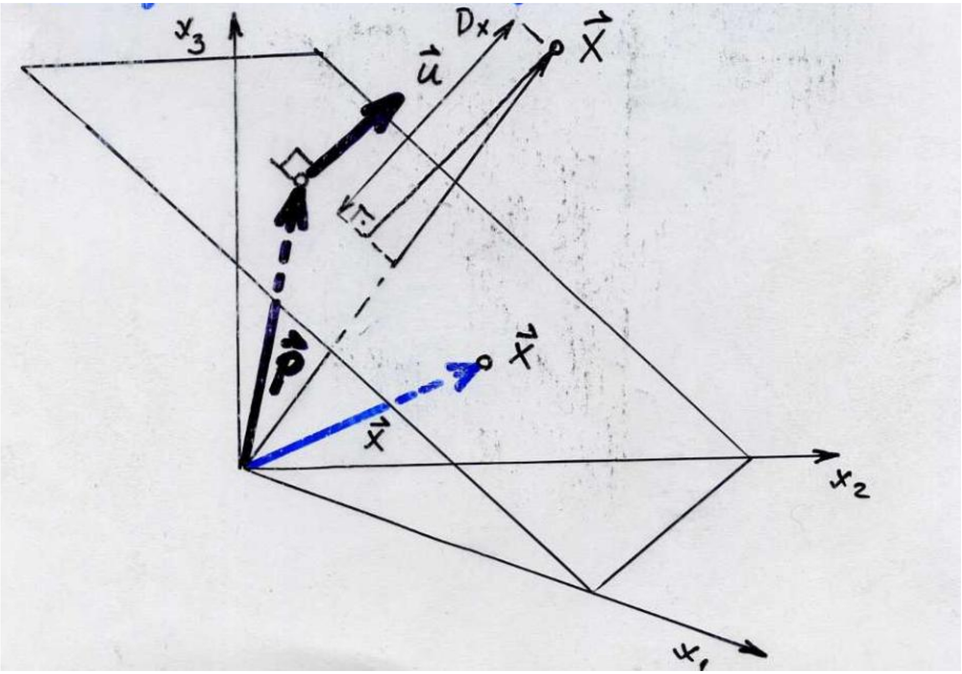
$$d(\mathbf{x}) = w_1x_1 + w_2x_2 + \dots + w_nx_n + w_{n+1} = 0$$

A decision function $d(\mathbf{x})$ is assumed to have the property:

$$\boxed{\begin{array}{c} d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} > 0 \text{ if } \mathbf{x} \in \omega_1 \text{ and} \\ d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} < 0 \text{ if } \mathbf{x} \in \omega_2 \end{array}}$$

If $d(\mathbf{x}) = 0$, $\mathbf{x}$ lies on hyperplane and its class membership is not defined

Geometrical interpretation of linear decision function



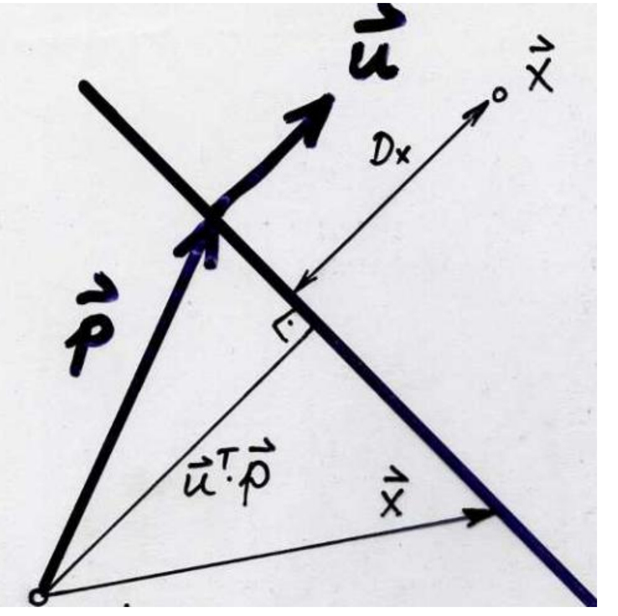
u – a unit vector normal to the hyperplane at some point **p** and oriented to the positive side of the hyperplane

Equation of the hyperplane:

$\mathbf{u}^T (\mathbf{x} - \mathbf{p}) = 0$

$\mathbf{u}^T \mathbf{x} = \mathbf{u}^T \mathbf{p}$

$\mathbf{u}^T \mathbf{x} = \mathbf{u}^T \mathbf{p}$ (1)



$d(\mathbf{x}) = \mathbf{w}_0^T \mathbf{x} + w_{n+1} = 0$

$\mathbf{w}_0^T \mathbf{x} = - w_{n+1} / \|\mathbf{w}_0\|$

$$\mathbf{w}_0^T \mathbf{x} / \|\mathbf{w}_0\| = - w_{n+1} / \|\mathbf{w}_0\| \qquad (2)$$

$$\|\mathbf{w}_0\| = \sqrt{w_1^2 + w_2^2 + \ldots + w_n^2}$$

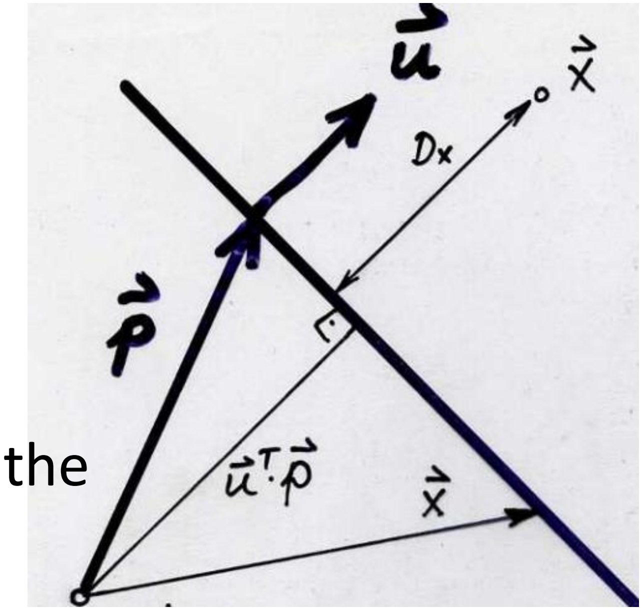
By comparing Eq. (1) and (2) follows:

$\mathbf{u} = \mathbf{w}_0 / \|\mathbf{w}_0\|$

$\mathbf{u}^T \mathbf{p} = - w_{n+1} / \|\mathbf{w}_0\|$

$\mathbf{u} = \mathbf{w}_0 / \|\mathbf{w}_0\|$

$\mathbf{u}^T \mathbf{p} = - w_{n+1} / \|\mathbf{w}_0\|$



origin

Absolute value of $\mathbf{u}^T \mathbf{p}$ represents the normal distance D_u from the origin to the hyperplane

$\mathbf{u} = \mathbf{w}_0 / \|\mathbf{w}_0\|$ defines the orientation of the hyperplane

If any component of **u** is zero, the hyperplane is parallel to the coordinate axis which corresponds to that component

- The normal distance D_x from hyperplane and an arbitrary pattern vector **x**:

$$D_x = \text{abs}(\mathbf{u}^T \mathbf{x} - \mathbf{u}^T \mathbf{p})$$

$$D_x = \text{abs}\left(\mathbf{w}_0^T \mathbf{x} / \|\mathbf{w}_0\| + w_{n+1} / \|\mathbf{w}_0\|\right)$$

Multicategory /multiclass/ case: M > 2

- there is more than one way to devise multicategory classifiers employing linear decision functions

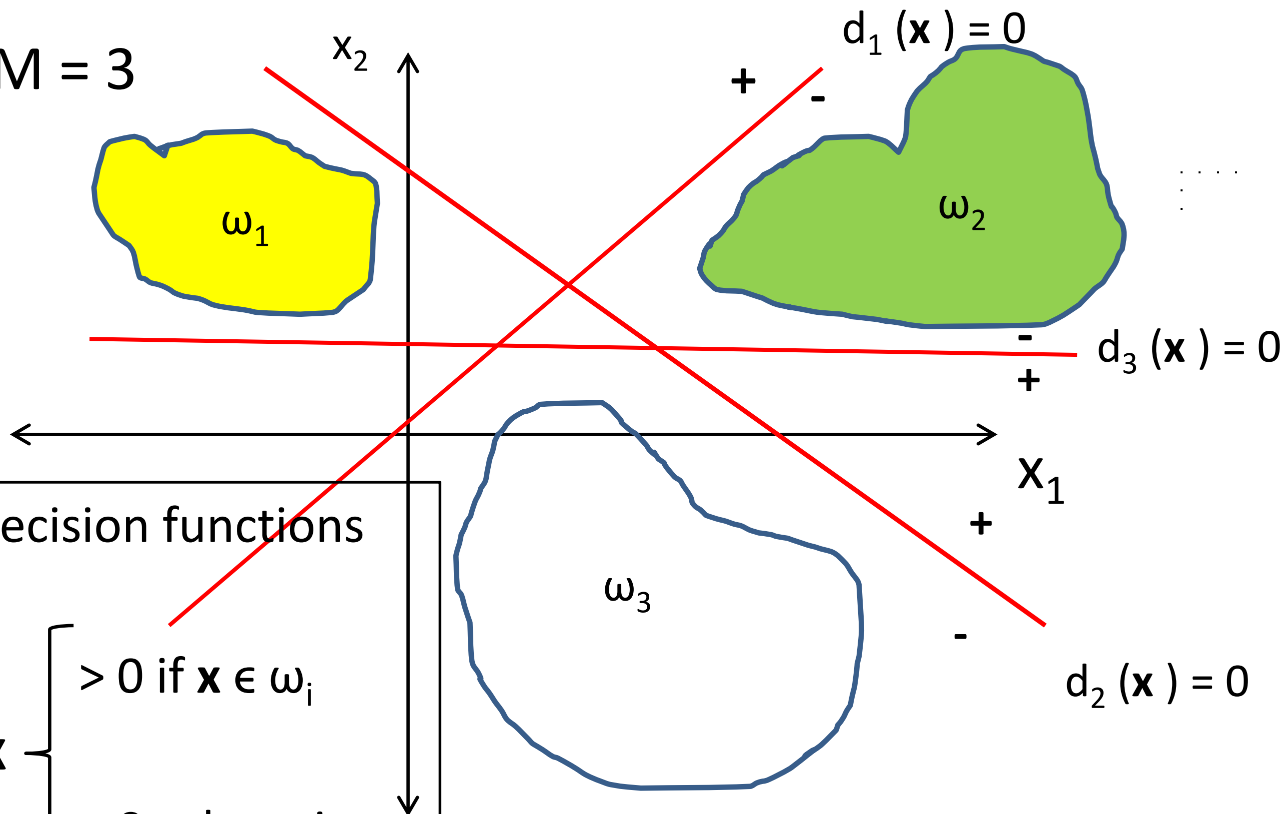
For example – reduce the problem to M two-classes problems where the *i*-th problem is solved by linear decision function that separates patterns assigned to ω_i from those not assigned to ω_i.

Three cases!

Case 1:

Each class is separable from other classes by a single decision surface.

Example: M = 3



There are M decision functions

$$d_i(\mathbf{x}) = \mathbf{w}^T \mathbf{x} \begin{cases} > 0 \text{ if } \mathbf{x} \in \omega_i \\ < 0 \text{ otherwise} \end{cases}$$

Case 2:

Each class is separable from every other individual class by a distinct decision surface – the classes are pairwise separable.

There are M(M-1)/2 decision surfaces;

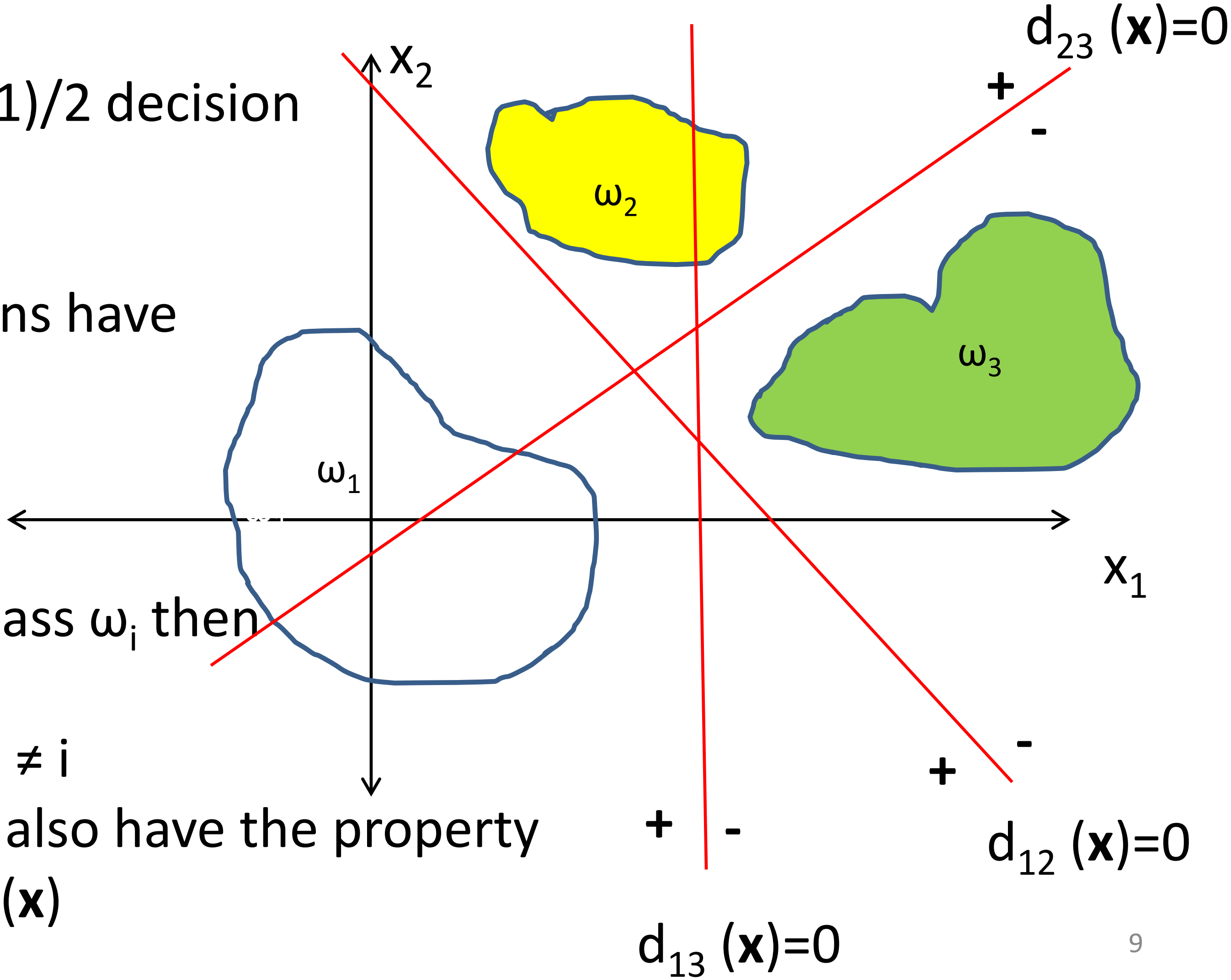
Decision functions have the form:

$d_{ij}(\mathbf{x}) = \mathbf{w}_{ij}^{T}\,\mathbf{x}$

If $\mathbf{x}$ belongs to class ω_i then

$d_{ij}(\mathbf{x}) > 0$ for all $j \neq i$

These functions also have the property that $d_{ij}(\mathbf{x}) = -\,d_{ji}(\mathbf{x})$



Case 3:

There exist M decision functions $d_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x}$, $k = 1, 2, \ldots, M$

If $\mathbf{x}$ belongs to class ω_i then

$d_i(\mathbf{x}) > d_j(\mathbf{x})$ for all $j \neq i$

Example:

$d_1(\mathbf{x}) = -x_1 + x_2$

$d_2(\mathbf{x}) = x_1 + x_2 - 1$

$d_3(\mathbf{x}) = -x_2$

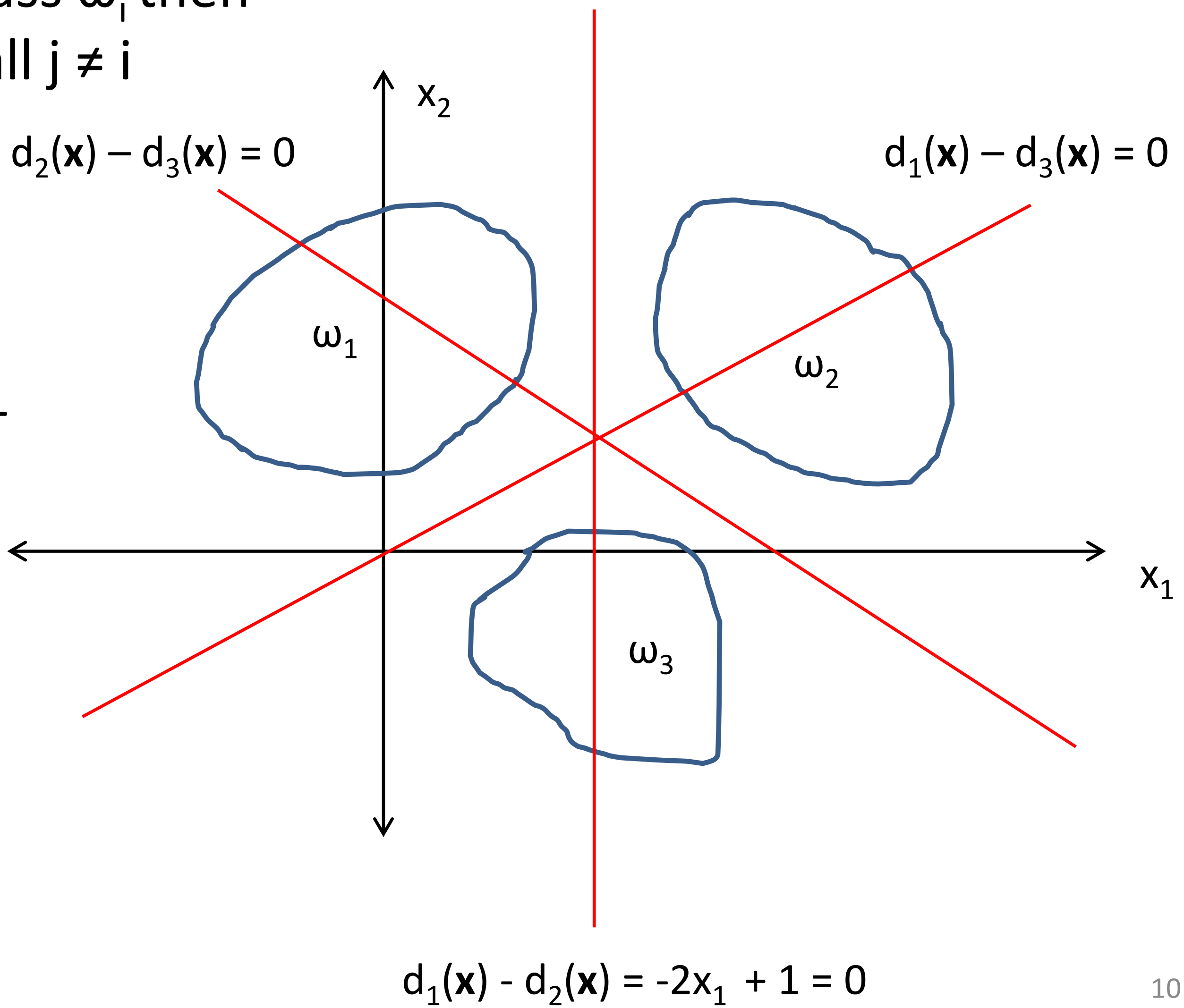
$\mathbf{x} = [2, 4]^T$

$d_1(\mathbf{x}) = 2$

$d_2(\mathbf{x}) = 5$

$d_3(\mathbf{x}) = -4$

$\mathbf{x} \in \omega_2$



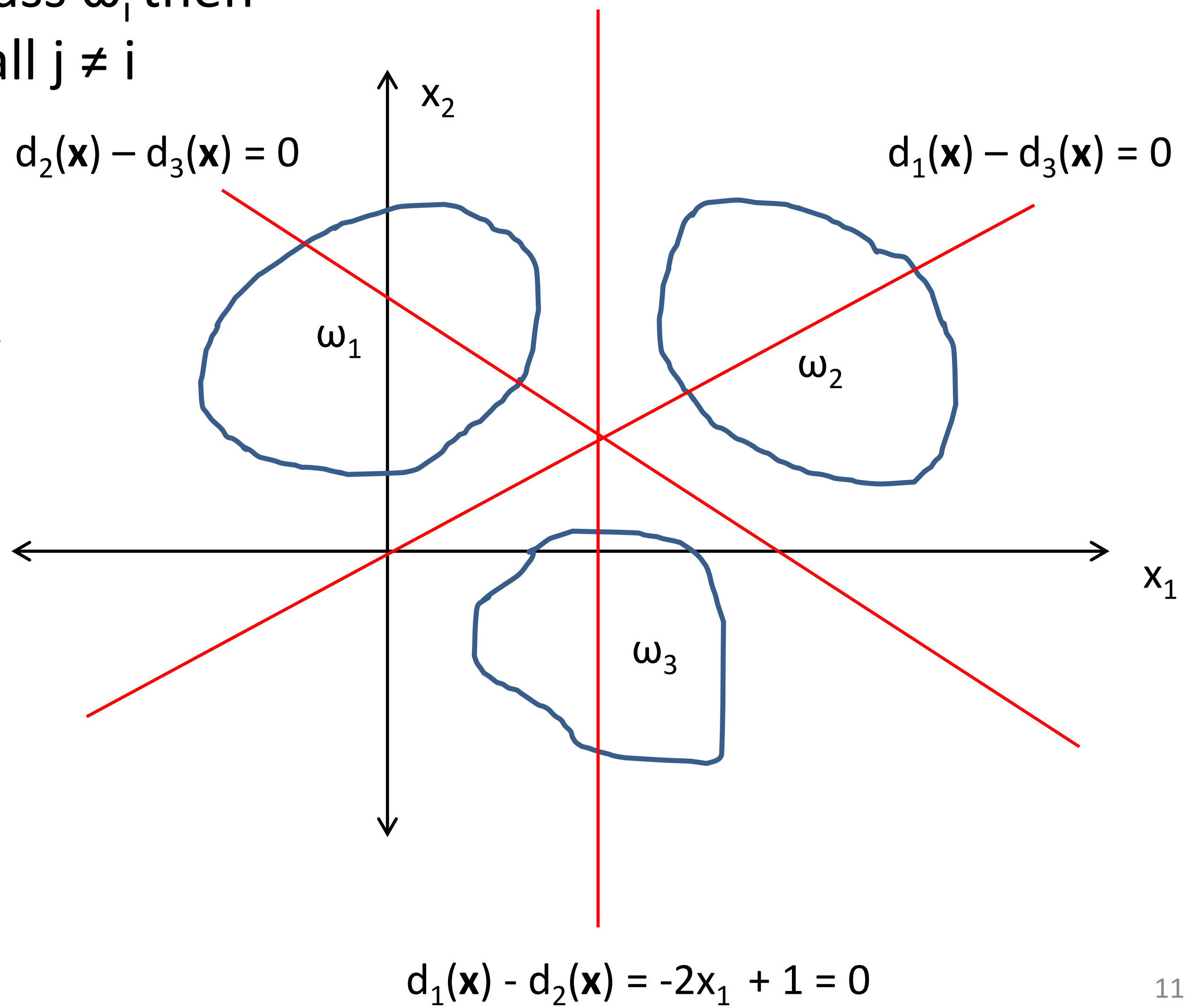
Case 3:

There exist M decision functions $d_k(\mathbf{x}) = \mathbf{w}_k^T\mathbf{x}$, $k = 1, 2, \ldots, M$

If $\mathbf{x}$ belongs to class ω_i then
$d_i(\mathbf{x}) > d_j(\mathbf{x})$ for all $j \neq i$

Case 3 is a special instance of Case 2 since we may define:

$$d_{ij}(\mathbf{x}) = d_i(\mathbf{x}) - d_j(\mathbf{x})$$
$$= (\mathbf{w}_i - \mathbf{w}_j)^T\mathbf{x}$$
$$= \mathbf{w}_{ij}^T\mathbf{x}$$



Learning algorithms for linear classifiers (deterministic approach)

Assumption: All feature vectors from available classes can be classified correctly using linear classifier

Advantages of linear classifiers: simplicity and computational attractiveness

$d(\mathbf{x}) = w_1x_1 + w_2x_2 + \ldots + w_nx_n + w_{n+1}$

$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$

Problem: determination of the components of the weight vector

$$\mathbf{w} = [w_1, w_2, \ldots , w_{n,} w_{n+1}]^T$$

Learning of decision function for M = 2 classes

$$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} > 0 \text{ if } \mathbf{x} \in \omega_1 \text{ and}$$
$$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} < 0 \text{ if } \mathbf{x} \in \omega_2$$

Learning of decision function for M = 2 classes (cont.)

A set of *N* training pattern vectors:

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ - labeled augmented pattern vectors

$\mathbf{x}_i = [x_1, x_2, \ldots, x_n, 1]^T; i = 1, 2, \ldots, N$

$$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} > 0 \text{ if } \mathbf{x} \in \omega_1 \text{ and}$$

$$d(\mathbf{x}) = \mathbf{w}^T \mathbf{x} < 0 \text{ if } \mathbf{x} \in \omega_2$$

If all pattern vectors of ω_2 are multiplied by -1 we obtain:

$$\mathbf{w}^T \mathbf{x} < 0 \;/\; (-1)$$

$$\mathbf{w}^T \mathbf{x} > 0$$

Now we have equivalent condition $\mathbf{w}^T \mathbf{x} > 0$ for all pattern vectors

Redefinition of the problem:

We are looking for the components of the weight vector $\mathbf{w} = [w_1, w_2, \ldots, w_n, w_{n+1}]^T$ which have to satisfy unique condition:

$$\mathbf{w}^T \mathbf{x} > 0$$

for all pattern vectors from training set.

We are looking for the components of the weight vector $\mathbf{w} = [w_1, w_2, \ldots, w_{n,} w_{n+1}]^T$ which have to satisfy unique condition:

$$\mathbf{w}^T \mathbf{x} > 0$$

for all pattern vectors from training set.

The above condition we can write in a form:

$$\mathbf{X}\,\mathbf{w} > \mathbf{0}, \text{ where}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^{\mathrm{T}} \\ \mathbf{x}_2^{\mathrm{T}} \\ \mathbf{x}_3^{\mathrm{T}} \\ \vdots \\ \mathbf{x}_N^{\mathrm{T}} \end{bmatrix}$$

and **0** is the zero vector.

Example:

$\mathbf{x}_1 = [0, 0]^T$, $\mathbf{x}_2 = [0, 1]^T$, $\mathbf{x}_3 = [1, 0]^T$, $\mathbf{x}_4 = [1, 1]^T$; $\mathbf{x}_1, \mathbf{x}_2 \in \omega_1$

$\mathbf{x}_3, \mathbf{x}_4 \in \omega_2$

Augmented pattern vectors are:

$\mathbf{x}_1 = [0, 0, 1]^T$, $\mathbf{x}_2 = [0, 1, 1]^T$, $\mathbf{x}_3 = [1, 0, 1]^T$, $\mathbf{x}_4 = [1, 1, 1]^T$

Multiply $\mathbf{x}_3$ and $\mathbf{x}_4$ by (-1) : $\mathbf{x}_3 = [-1, 0, -1]^T$ and $\mathbf{x}_4 = [-1, -1, -1]^T$

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} > \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The weight vector **w** which satisfies the system of inequalities

$$\mathbf{X}\,\mathbf{w} > \mathbf{0}$$

is separating vector.

If **w** exists – classes ω_1 and ω_2 are (linear) separable.

Note: All pattern vectors from ω_2 are multiplied by -1.

The gradient technique

In general, the gradient of function f(**y**) with respect to the vector **y** = $[y_1, y_2, \dots, y_n]^T$ is defined as:

$$\operatorname{grad} f(\mathbf{y}) = \frac{df(\boldsymbol{y})}{d\boldsymbol{y}} = \begin{pmatrix} \frac{df}{dy_1} \\ \frac{dy}{dy_2} \\ \cdot \\ \cdot \\ \cdot \\ \frac{df}{dy_n} \end{pmatrix}$$

-grad $f(\mathbf{y})$: Gradient of a scalar function of a vector argument is a vector
- each component of the gradient gives the rate of change of the function in the direction of that component

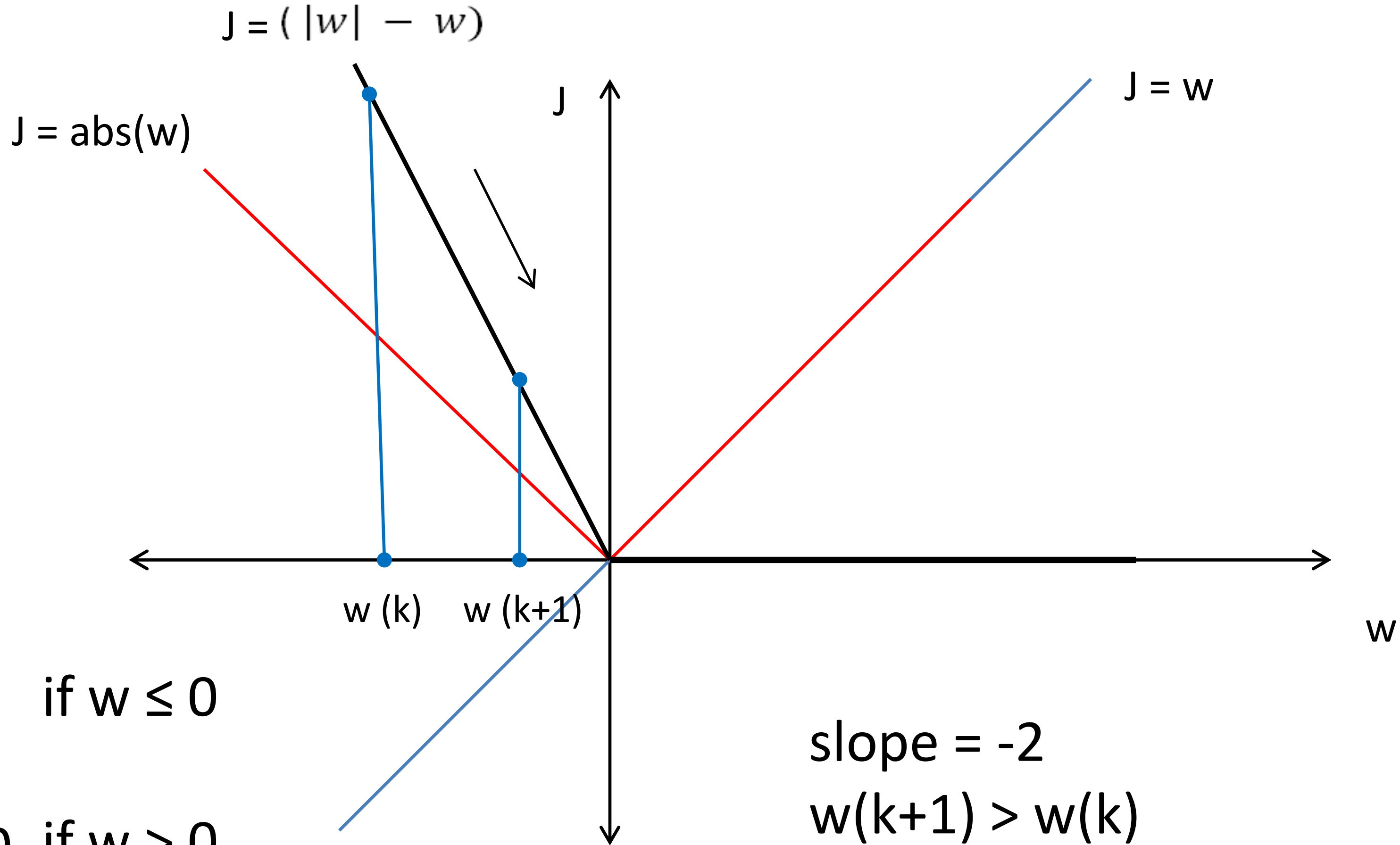
- the positive of gradient points in direction of the maximum rate of increase of the function f when the argument increases

- the negative of the gradient points in the direction of the maximum rate of decrease of the function f

- the above properties can be used for finding the minimum (or maximum) of a function f

Example:

Consider the function J(w, 1) = $(|w| - w)$



$$\frac{\partial J}{\partial w} = -2 \quad \text{if } w \leq 0$$

$$\frac{\partial J}{\partial w} = 0 \quad \text{if } w > 0$$

slope = -2

w(k+1) > w(k)

we are approaching

to the minimum of J

Basic idea - select the proper function J(**w**, **x**) /criterion function/ which achieves the minimum when the condition $\mathbf{w}^T \mathbf{x}_i > 0$ for all $\mathbf{x}_i$, i = 1, 2, … , N, is satisfied!

- Incrementing **w** in the direction of negative gradient of J(**w**, **x**), in order to seek the minimum of the function.

Gradient decent algorithm

w(k) – value of **w** at the *k*th step

$$\mathbf{w}(k+1) = \mathbf{w}(k) - c \left(\frac{\partial J(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}}\right)_{\mathbf{w} = \mathbf{w}(k)}$$

where **w**(k + 1) represents the new value of **w**, and c > 0 dictates the magnitude of the correction.

Perceptron algorithm

Let us define the criterion function:

$$J(\mathbf{w}, \mathbf{x}) = \frac{1}{2} \left(|\mathbf{w}^T \mathbf{x}| - \mathbf{w}^T \mathbf{x} \right)$$

The partial derivation of J with respect to **w**:

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, \mathbf{x}) = \frac{1}{2} \left(\mathbf{x} \operatorname{sgn}(\mathbf{w}^T \mathbf{x}) - \mathbf{x} \right)$$

where:

$$\operatorname{sgn}(x) \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x \leq 0 \end{cases}$$

Some rules for partial derivations:

$$\frac{d\,(\mathbf{x}^T\,\mathbf{A})}{d\mathbf{x}} = \mathbf{A}$$

$$\frac{d}{d\mathbf{x}}\,(\mathbf{x}^T) = \mathbf{I}$$

$$\frac{d}{d\mathbf{x}}\,(\mathbf{x}^T\mathbf{a}) = \frac{d}{d\mathbf{x}}\,(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$$

$$\frac{d}{d\mathbf{X}}\,(\mathbf{a}^T X\,\mathbf{b}) = \mathbf{a}\,\mathbf{b}^T$$

$$\frac{d}{d\mathbf{X}}(\mathbf{a}^T X\,\mathbf{a}) = \frac{d}{d\mathbf{X}}(\mathbf{a}^T X^T\mathbf{a}) = \mathbf{a}\,\mathbf{a}^T$$

$$\frac{d}{d\mathbf{x}}\,(\mathbf{x}^T\,\mathrm{C}\,\mathbf{x}) = (\mathrm{C} + C^T)\,x^T$$

$$\text{if } \mathrm{C} = \mathrm{C}^{\mathrm{T}} \text{ then } \frac{d}{d\mathbf{x}}\,(x^T\,\mathbf{C}\,x) = 2\,\mathbf{Cx}$$

$$\frac{d}{d\mathbf{x}}\,(\mathbf{x}^T\mathbf{x}) = 2\,\mathbf{x}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{Ax} + \mathbf{b})^T(\mathbf{Dx} + \mathbf{e}) = A^T(\mathbf{Dx} + \mathbf{e}) + D^T(A\mathbf{x} + \mathbf{b})$$

$$\operatorname{sgn}(\mathbf{w}^T\mathbf{x}) \begin{cases} 1 & \text{if } \mathbf{w}^T\mathbf{x} > 0 \\ -1 & \text{if } \mathbf{w}^T\mathbf{x} \leq 0 \end{cases}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^T\mathbf{a}) = \frac{d}{d\mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$$

$$\frac{\partial}{\partial \mathbf{w}} J(\mathbf{w}, \mathbf{x}) = \frac{1}{2}\left(\mathbf{x}\operatorname{sgn}(\mathbf{w}^T\mathbf{x}) - \mathbf{x}\right) \qquad (1)$$

Substituting Eq. (1) into $\mathbf{w}(k+1) = \mathbf{w}(k) - c\left(\frac{\partial J(\boldsymbol{w},\boldsymbol{x})}{\partial \boldsymbol{w}}\right)_{\mathbf{w}=\mathbf{w}(k)}$

yields:

$$\mathbf{w}(k+1) = \mathbf{w}(k) - c\left(\frac{1}{2}\left(\mathbf{x}(k)\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}(k)\right) - \mathbf{x}(k)\right)\right)$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \frac{c}{2}\left(\mathbf{x}(k) - \mathbf{x}(k)\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}(k)\right)\right)$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) - c\left(\frac{1}{2}\left(\mathbf{x}(k)\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}(k)\right) - \mathbf{x}(k)\right)\right)$$

$$\mathbf{w}(k+1) = \mathbf{w}(k) + \frac{c}{2}\left[\mathbf{x}(k) - \mathbf{x}(k)\operatorname{sgn}\left(\mathbf{w}^T\mathbf{x}(k)\right)\right]$$

$$\boxed{\mathbf{w}(k+1) = \mathbf{w}(k) + c\begin{cases}\mathbf{0} & \text{if } \boldsymbol{w}^T(k)\mathbf{x}(k) > 0\\ \mathbf{x}(k) & \text{if } \mathbf{w}^T(k)\mathbf{x}(k) \le 0\end{cases}}$$

where $c > 0$ and $\mathbf{w}(1)$ is arbitrary.

$$\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}$$

The criterion function:

$$J(\mathbf{w},\mathbf{x}) = \frac{1}{4\mathbf{x}^{\mathrm{T}}\mathbf{x}}\left(\left|\mathbf{w}^{\mathrm{T}}\mathbf{x}\right|^2 - \underline{\left|\mathbf{w}^{\mathrm{T}}\mathbf{x}\right|\mathbf{w}^{T}\mathbf{x}}\right)$$

The partial derivative of above function:

$$\frac{\partial J(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}} = \frac{1}{4\mathbf{x}^{T}\mathbf{x}}[2|\mathbf{w}^{T}\mathbf{x}|\mathbf{x}\,\mathrm{sgn}(\mathbf{w}^{T}\mathbf{x}) - (|\mathbf{w}^{T}\mathbf{x}|\mathbf{x} + (\mathbf{w}^{T}\mathbf{x})\mathbf{x}\,\mathrm{sgn}(\mathbf{w}^{T}\mathbf{x})]$$

$$(\mathbf{w}^{T}\mathbf{x})\mathbf{x}\,\mathrm{sgn}(\mathbf{w}^{T}\mathbf{x}) = \left|\mathbf{w}^{\mathrm{T}}\mathbf{x}\right|\mathbf{x}$$

$$\frac{\partial J(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}} = \frac{1}{4\mathbf{x}^T\mathbf{x}}[2|\mathbf{w}^T\ \mathbf{x}|\mathbf{x}\ \mathrm{sgn}(\mathbf{w}^T\mathbf{x}) - 2\ |\mathbf{w}^T\ \mathbf{x}|\mathbf{x}]$$

$$\frac{\partial J(\mathbf{w},\mathbf{x})}{\partial \mathbf{w}} = \frac{1}{2\mathbf{x}^T\mathbf{x}}[|\mathbf{w}^T\ \mathbf{x}|\mathbf{x}\ \mathrm{sgn}(\mathbf{w}^T\mathbf{x}) - \ |\mathbf{w}^T\ \mathbf{x}|\mathbf{x}]$$

$$\mathbf{w}(k+1) = \ \mathbf{w}(k) + \lambda\frac{|\mathbf{w}^T(k)\ \mathbf{x}(k)|}{2\mathbf{x}(k)^T\mathbf{x}(k)}[\boldsymbol{x}(k) - \ \mathbf{x}(k)\ \mathrm{sgn}(\mathbf{w}(k)^T\mathbf{x}(k))]$$

The fraction-correction algorithm:

$$\boxed{\mathbf{w}(k+1) = \mathbf{w}(k) + \lambda\frac{|\mathbf{w}^T(k)\ \mathbf{x}(k)|}{\mathbf{x}(k)^T\mathbf{x}(k)}\begin{cases}\mathbf{0}\ \text{if} & \mathbf{w}^T(k)\mathbf{x}(k) > 0\\ \mathbf{x}(k)\ \text{if} & \mathbf{w}^T(k)\mathbf{x}(k) \leq 0\end{cases}}$$

λ – correction factor; $0 < \lambda < 2$

$\mathbf{w}(1) \neq \mathbf{0}$

Variations of perceptron algorithm

$$\mathbf{w}(k+1) = \mathbf{w}(k) + c\begin{cases} \mathbf{0} & \text{if } \boldsymbol{w}^{T}(k)\mathbf{x}(k) > 0 \\ \mathbf{x}(k) & \text{if } \mathbf{w}^{T}(k)\mathbf{x}(k) \le 0 \end{cases}$$

i) Fixed-increment algorithm
ii) Algorithm with absolute correction
iii) Fractional-correction algorithm

i) $c > 0$
ii) If $\mathbf{w}(k)^{T}\mathbf{x}(k) \le 0$ then select c that:

$$\boldsymbol{w}^{T}(k+1)\,\mathbf{x}(k) = [\mathbf{w}(k) + c\mathbf{x}(k)]^{T}\mathbf{x}(k) > 0$$

integer value $c > \dfrac{|\mathbf{w}^{T}(k)\,\mathbf{x}(k)|}{\mathbf{x}^{T}(k)\mathbf{x}(k)}$

iii)

$$\left|\mathbf{w}^{T}(\mathrm{k})\,\mathbf{x}(\mathrm{k}) - \ \mathbf{w}^{T}(\mathrm{k}+1)\mathbf{x}(\mathrm{k})\right| = \lambda\left|\mathbf{w}^{T}(\mathrm{k})\,\mathbf{x}(\mathrm{k})\right|$$

$$\mathbf{w}(\mathrm{k}\ +1)\ =\ \mathbf{w}(\mathrm{k})\ +\ \mathrm{c}\,\mathbf{x}(\mathrm{k})$$

$$\left|\mathbf{w}^{T}(\mathrm{k})\,\mathbf{x}(\mathrm{k}) - \ (\mathbf{w}^{T}(\mathrm{k}) + \mathrm{c}\,\mathbf{x}^{\mathbf{T}}(\mathrm{k}))\,\mathbf{x}(\mathrm{k})\right| = \lambda\left|\mathbf{w}^{T}(\mathrm{k})\,\mathbf{x}(\mathrm{k})\right|$$

$$\mathrm{c} = \lambda\frac{\left|\mathbf{w}^{T}(\mathrm{k})\,\mathbf{x}(\mathrm{k})\right|}{\mathbf{x}^{T}(\mathrm{k})\mathbf{x}(\mathrm{k})}$$

Example:

-Training set: $\omega_1 = \{(0, 0)^T, (0, 1)^T\}$

$\omega_2 = \{(1, 0)^T, (1, 1)^T\}$

c =1

Augmented pattern vectors:

$\omega_1 = \{(0, 0, 1)^T, (0, 1, 1)^T\}$

$\omega_2 = \{(1, 0, 1)^T, (1, 1, 1)^T\}$

Pattern vectors from ω_2 have to be multiplied by (-1)

$\omega_2 = \{(-1, 0, -1)^T, (-1, -1, -1)^T\}$

Let us suppose that vectors are linear separable by d(**x**):

$d(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = w_1x_1 + w_2 x_2 + w_3$

Let us select $\mathbf{w}(1) = (-1, 0, 0)^T$

1.Step

$\mathbf{w}^T(1)\,\mathbf{x}(1) = (-1, 0, 0)\begin{pmatrix}0\\0\\1\end{pmatrix} = 0,$

$\mathbf{w}(2) = \mathbf{w}(1) + \mathbf{x}(1) = \begin{pmatrix}-1\\0\\1\end{pmatrix}$

2. Step

$\mathbf{w}^T(2)\,\mathbf{x}(2) = (-1, 0, 1)\begin{pmatrix}0\\1\\1\end{pmatrix} = 1$

$\mathbf{w}^T(2)\,\mathbf{x}(2) > 0$ .... $\mathbf{w}(3) = \mathbf{w}(2)$

3. Step

$\mathbf{w}^T(3)\,\mathbf{x}(3) = (-1, 0, 1)\begin{pmatrix}-1\\0\\-1\end{pmatrix} = 0$

3. Step (cont.)

$\mathbf{w}(4) = \mathbf{w}(3) + \mathbf{x}(3)$

$$\mathbf{w}(4) = \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \\ -1 \end{pmatrix} = \begin{pmatrix} -2 \\ 0 \\ 0 \end{pmatrix}$$

4. Step

$$\mathbf{w}^T(4)\,\mathbf{x}(4) = (-2, 0, 0) \begin{pmatrix} -1 \\ -1 \\ -1 \end{pmatrix} = 2$$

$\mathbf{w}(5) = \mathbf{w}(4)$

There were corrections on the weight vector in the 1. and 3. steps!

Solution has been obtained ONLY when algorithm yields a complete ERROR-FREE iteration through all patterns

The second iteration through all pattern vectors:

$x(5) = x(1)$; $x(6) = x(2)$; $x(7) = x(3)$; $x(8) = x(4)$

5. Step

$$\mathbf{w}^T(5)\,\mathbf{x}(5) = (-2, 0, 0)\begin{pmatrix}0\\0\\1\end{pmatrix} = 0$$

$$\mathbf{w}(6) = \mathbf{w}(5) + \mathbf{x}(5) = \begin{pmatrix}-2\\0\\0\end{pmatrix} + \begin{pmatrix}0\\0\\1\end{pmatrix} = \begin{pmatrix}-2\\0\\1\end{pmatrix}$$

6. Step

$$\mathbf{w}^T(6)\,\mathbf{x}(6) = (-2, 0, 1)\begin{pmatrix}0\\1\\1\end{pmatrix} = 1$$

$$\mathbf{w}(7) = \mathbf{w}(6)$$

7. Step

$$\mathbf{w}^T(7)\,\mathbf{x}(7) = (-2, 0, 1)\begin{pmatrix}-1\\0\\-1\end{pmatrix} = 1$$

$$\mathbf{w}(8) = \mathbf{w}(7)$$

8. Step

$$\mathbf{w}^T(8)\,\mathbf{x}(8) = (-2, 0, 1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix} = 1$$

Two error occurred in this iteration!

$x(9) = x(1)$; $x(10) = x(2)$; $x(11) = x(3)$; $x(12) = x(4)$

Third iteration trough all vectors

x(9) = x(1); x(10) = x(2); x(11) = x(3); x(12) = x(4)

9. Step

$\mathbf{w}^T(9)\,\mathbf{x}(9) = (-2, 0, 1)\begin{pmatrix}0\\0\\1\end{pmatrix} = 1$

$\mathbf{w}(10) = \mathbf{w}(9)$

10. Step

$\mathbf{w}^T(10)\,\mathbf{x}(10) = (-2,0, 1)\begin{pmatrix}0\\1\\1\end{pmatrix} =1$

$\mathbf{w}(11) = \mathbf{w}(10)$

11. Step

$\mathbf{w}^T(11)\,\mathbf{x}(11) = (-2, 0, 1)\begin{pmatrix}-1\\0\\-1\end{pmatrix} = 1$

$\mathbf{w}(12) = \mathbf{w}(11)$

12. Step

$\mathbf{w}^T(12)\,\mathbf{x}(12) = (-2, 0, 1)\begin{pmatrix}-1\\-1\\-1\end{pmatrix} = 1$

All pattern vectors are classified correctly - Error-free

Solution:

$\mathbf{w} = (-2, 0, 1)^T$

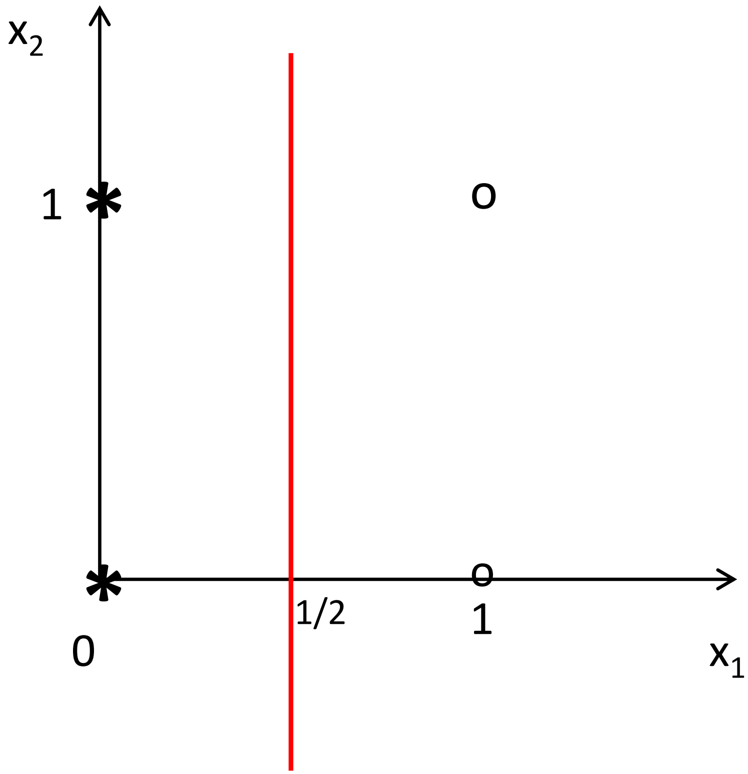
Decision function:

$d(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = -2x_1 + 1 = 0$

$x_1 = 1/2$

$*\ \in \omega_1$

$o\ \in \omega_2$



Generalized perceptron algorithm
(Multicategory classification)
Training method for M > 2
Case 3:
$\omega_1, \omega_2, \ldots, \omega_M$

If $\mathbf{x} \in \omega_i$ then $d_i(\mathbf{x}) > d_j(\mathbf{x})$ for all $j \neq i$;
At *k*-th iterative step during training, a pattern vector $\mathbf{x}(k)$ belonging to class ω_i is presented to the machine;
Evaluate M decision functions
$d_j(\mathbf{x}(k)) = \mathbf{w}_j^T \mathbf{x}(k)$, j = 1, 2, ... , M
Then, if
$$d_i(\mathbf{x}(k)) > d_j(\mathbf{x}(k)) \qquad j = 1, 2, \ldots, M;\ j \neq i;$$
the weight vectors are not adjusted:
$$\mathbf{w}_j(k+1) = \mathbf{w}_j(k) \qquad j = 1, 2, \ldots, M$$

If $d_i(\mathbf{x}(k)) \leq d_L(\mathbf{x}(k))$

The following weight adjustments are made:

$$\mathbf{w}_i(k + 1) = \mathbf{w}_i(k) + c\mathbf{x}(k)$$

$$\mathbf{w}_L(k + 1) = \mathbf{w}_L(k) - c\mathbf{x}(k)$$

$$\mathbf{w}_j(k + 1) = \mathbf{w}_j(k) \text{ for } j = 1, 2, \ldots, M;\ j \neq i;\ j \neq L$$

c is positive constant.

If the classes are separable under *Case 3,* the algorithm converges in the finite number of iterations.

Example:

M = 3

Each class contains a single pattern:

$\omega_1 = \{(0, 0)^T\}$

$\omega_2 = \{(1, 1)^T\}$

$\omega_3 = \{(-1, 1)^T\}$

0. Augment the pattern vectors:

$\omega_1 = \{(0, 0, 1)^T\}$

$\omega_2 = \{(1, 1, 1)^T\}$

$\omega_3 = \{(-1, 1, 1)^T\}$

Note that none of pattern vectors is multiplied by -1.

$c = 1$; $\mathbf{w}_1(1) = \mathbf{w}_2(1) = \mathbf{w}_3(1) = (0, 0, 0)^T$

Apply the **Generalized perceptron algorithm!**

- Find subspaces of feature space which correspond to each class!

Solution:

$d_1(\mathbf{x}) = -2x_2$

$d_2(\mathbf{x}) = 2x_1 - 2$

$d_3(\mathbf{x}) = -2x_1 - 2$

Let us check the solution:

$\mathbf{x} = (0, 0)^T \in \omega_1$

$d_1(\mathbf{x}) = 0$

$d_2(\mathbf{x}) = -2$

$d_3(\mathbf{x}) = -2$

$\mathbf{x} = (1, 1)^T \in \omega_2$

$d_1(\mathbf{x}) = -2$

$d_2(\mathbf{x}) = 0$

$d_3(\mathbf{x}) = -4$

$\mathbf{x} = (-1, 1)^T \in \omega_3$

$d_1(\mathbf{x}) = -2$

$d_2(\mathbf{x}) = -4$

$d_3(\mathbf{x}) = 0$

LMSE – Least-Mean-Square-Error Algorithm
Ho-Kashyap algorithm

-The perceptron algorithm and its variations converge when the classes are separable by linear decision functions

- In non-separable situations these algorithms oscillate
- It is not possible to precompute the number of steps required for convergence in linear separable situation
- Ho-Kashyap algorithm indicates that the classes are not separable!

Instead of finding weight vector **w** such that **Xw** > **0** is satisfied, Ho-Kashyap algorithm searches for vectors **w** and **b** such that:

$$\mathbf{X}\,\mathbf{w} = \mathbf{b}$$

where the components of **b** = $(b_1, b_2, \dots, b_N)^T$ are all positive.
N – number of pattern vectors in training set
The criterion function:

$$J(\mathbf{w}, \mathbf{x}, \mathbf{b}) = \frac{1}{2}\sum_{j=1}^{N}(\mathbf{w}^T\mathbf{x}_j - b_j)^2 = \frac{1}{2}\|\mathbf{Xw} - \mathbf{b}\|^2 = \frac{1}{2}(\mathbf{Xw} - \mathbf{b})^T(\mathbf{Xw} - \mathbf{b})$$

where $\|\mathbf{Xw} - \mathbf{b}\|$ is the magnitude of the vector ($\mathbf{Xw} - \mathbf{b}$).
The function J(**w**, **x**, **b**) achieves its minimum whenever

$$\mathbf{X}\,\mathbf{w} = \mathbf{b}$$

- Both variables **w** and **b** can be used in the minimization procedure;
We expect that above can improve the convergence rate of the algorithm;
-Function J(**w**, **x**, **b**) will be minimized with respect to **w** and **b**;

- Gradients:

$$\frac{\partial \mathrm{J}}{\partial \mathbf{w}} \text{ and } \frac{\partial \mathrm{J}}{\partial \mathbf{b}}$$

$$\boxed{\frac{\partial \mathrm{J}}{\partial \mathbf{w}}}$$

$$\mathrm{J}(\mathbf{w}, \mathbf{x}, \mathbf{b}) = \frac{1}{2}\,(\mathbf{Xw} - \mathbf{b})^{\mathrm{T}}(\mathbf{Xw} - \mathbf{b})$$

$$\frac{d}{d\mathbf{x}}\,(\mathbf{x}^{T}\mathbf{x}) \;=\; 2\,\mathbf{x}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{Ax} + \mathbf{b})^{T}(\mathbf{Dx} + \mathbf{e}) = A^{T}(\mathbf{Dx} + \mathbf{e}) + D^{T}(A\mathbf{x} + \mathbf{b})$$

$$J(\mathbf{w}, \mathrm{x}, \mathbf{b}) = \frac{1}{2}\,(\mathbf{Xw} - \mathbf{b})^{\boldsymbol{T}}(\mathbf{Xw} - \mathbf{b})$$

$$\frac{\mathrm{d}}{\mathrm{d}\mathbf{x}}(\mathbf{Ax} + \mathbf{b})^{\mathrm{T}}(\mathbf{Dx} + \mathbf{e}) = \mathbf{A}^{\mathrm{T}}(\mathbf{Dx} + \mathbf{e}) + \mathbf{D}^{\mathrm{T}}(\mathbf{Ax} + \mathbf{b})$$

A = **X**, **e = - b**, **D** = **X**, **b = - b**

$$\frac{\partial J}{\partial \mathbf{w}} = \frac{\mathbf{1}}{\mathbf{2}}\left(\mathbf{X}^{\mathrm{T}}(\mathbf{Xw} - \mathbf{b}) + \mathbf{X}^{\mathrm{T}}(\mathbf{Xw} - \mathbf{b})\right)$$

$$\boxed{\frac{\partial J}{\partial \mathbf{w}} = \mathbf{X}^{\mathrm{T}}(\mathbf{Xw} - \mathbf{b})}$$

$$\boxed{\frac{\partial \mathrm{J}}{\partial \mathbf{b}}}$$

$$\mathrm{J}(\mathbf{w}, \mathbf{x}, \mathbf{b}) = \frac{1}{2}(\mathbf{Xw} - \mathbf{b})^{T}(\mathbf{Xw} - \mathbf{b})$$

$$\frac{\partial \mathrm{J}(\mathbf{w}, \mathbf{x}, \mathbf{b})}{\partial \mathbf{b}} = \frac{\partial \mathrm{J}(\mathbf{w}, \mathbf{x}, \mathbf{b})}{\partial \mathbf{b}}\left\{\frac{1}{2}\left[(X\,\mathbf{w})^{\mathrm{T}}\mathbf{Xw} - \boldsymbol{b}^{\mathrm{T}}\mathbf{Xw} - (\mathbf{X}\,\mathbf{w})^{\mathrm{T}}\mathbf{b} + \mathbf{b}^{\mathrm{T}}\mathbf{b}\right]\right\}$$

$$\frac{d}{d\mathbf{x}}(\mathbf{x}^{T}\mathbf{a}) = \frac{d}{d\mathbf{x}}(\mathbf{a}^{T}\mathbf{x}) = \mathbf{a}$$

$$\frac{\partial \mathrm{J}(\mathbf{w}, \mathbf{x}, \mathbf{b})}{\partial \mathbf{b}} = \frac{1}{2}[-\mathbf{Xw} - \mathbf{Xw} + 2\mathbf{b}]$$

$$\boxed{\frac{\partial}{\partial \mathbf{b}}\mathrm{J}(\mathbf{w}, \mathbf{x}, \mathbf{b}) = -(\mathbf{Xw} - \mathbf{b})}$$

Since **w** is not constrained in any way, and we can set

$$\frac{\partial J}{\partial \mathbf{w}} = \mathbf{0}$$

$$\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{b}) = \mathbf{0}$$

$$X^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{b}$$

$$X^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{b} \;\; / \; (X^T\mathbf{X})^{-1}$$

$$\mathbf{w} = (X^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{b}$$

$(X^T\mathbf{X})^{-1}\,\mathbf{X}^T = X^{\#}$ - generalized inverse matrix

$$\boxed{\mathbf{w} = \mathbf{X}^{\#}\,\mathbf{b}} \longrightarrow \mathbf{w}(k+1) = \mathbf{X}^{\#}\,\mathbf{b}\,(k+1)$$

- A vector **b** = $(b_1, b_2, \ldots, b_n)^T$ is a positive vector – all components of **b** are constrained to be positive
-This vector must be varied in a such a manner as never to violate this constraint:

$$\mathbf{b}(\mathrm{k}+1) = \mathbf{b}(\mathrm{k}) + \delta\mathbf{b}(\mathrm{k})$$

(*)

$$\delta b_i(\mathrm{k}) = \begin{cases} 2\mathrm{c}[\mathbf{X}\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k})]_{\mathrm{i}} & \text{if } [\mathbf{X}\,\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k})]_{\mathrm{i}} > 0 \\ 0 & \text{if } [\mathbf{X}\,\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k})]_{\mathrm{i}} \le 0 \end{cases}$$

k denotes the iteration index, *i* denotes the index of the vector components, *c* is positive correction increment
- Equation (*) may be written in vector form:

$$\delta\mathbf{b}(\mathrm{k}) = \mathrm{c}[\mathbf{X}\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k}) + |\mathbf{X}\mathbf{w}(\mathrm{k}) - \boldsymbol{b}(k)|]$$

- where $|\mathbf{X}\mathbf{w}(\mathrm{k}) - \boldsymbol{b}(k)|$ indicates the absolute value of each component of the vector $[\mathbf{X}\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k})]$

$$\delta\mathbf{b}(\mathrm{k}) = c[\mathbf{X}\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k}) + |\mathbf{X}\mathbf{w}(\mathrm{k}) - \boldsymbol{b}(k)|]$$

$$\mathbf{w} = \mathbf{X}^{\#}\,\mathbf{b}$$

$$\mathbf{w}(\mathrm{k}+1) = \mathbf{X}^{\#}\,\mathbf{b}\,\underline{(\mathrm{k}+1)}$$

$$\mathbf{w}(\mathrm{k}+1) = \mathbf{X}^{\#}\underline{[\mathbf{b}(\mathrm{k}) + \delta\mathbf{b}\,(\mathrm{k})]}$$

$$\mathbf{w} = \mathbf{X}^{\#}\,\mathbf{b}$$

$$\mathbf{w}(\mathrm{k}+1) = \mathbf{X}^{\#}\mathbf{b}(\mathrm{k}) + \mathbf{X}^{\#}\delta\mathbf{b}\,(\mathrm{k})$$

$$\mathbf{w}(\mathrm{k}+1) = \mathbf{w}(\mathrm{k}) + \mathbf{X}^{\#}\delta\mathbf{b}\,(\mathrm{k})$$

- let us denote

$$\mathbf{e}(\mathrm{k}) = \mathbf{X}\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k})$$

- we have the following algorithm:

$$\mathbf{w}(1) = \mathbf{X}^{\#}\mathbf{b}(1) \ , \qquad b_i > 0$$

$$\mathbf{e}(k) = \mathbf{X}\mathbf{w}(k) - \mathbf{b}(k)$$

$$\delta\mathbf{b}(\mathrm{k}) = \mathrm{c}[\mathbf{X}\mathbf{w}(\mathrm{k}) - \mathbf{b}(\mathrm{k}) + |\mathbf{X}\mathbf{w}(\mathrm{k}) - \boldsymbol{b}(k)|]$$

$$\delta\mathbf{b}(\mathrm{k}) = \mathrm{c}[\mathbf{e}(\mathrm{k}) + |\mathbf{e}(\mathrm{k})|]$$

Where $|\mathbf{e}(\mathrm{k})|$ denotes the vector whose components are the absolute value of the component **e**(k)

$$\mathbf{w}(\mathrm{k}+1) = \mathbf{w}(\mathrm{k}) + \mathrm{c}\mathbf{X}^{\#}[\mathbf{e}(\mathrm{k}) + |\mathbf{e}(\mathrm{k})|]$$

$$\mathbf{b}(\mathrm{k}+1) = \mathbf{b}(\mathrm{k}) + \mathrm{c}[\mathbf{e}(\mathrm{k}) + |\mathbf{e}(\mathrm{k})|]$$

- **w**(k+1) can be also calculated:

$$\mathbf{w}(k+1) = \mathbf{X}^{\#}\mathbf{b}(k+1)$$

Algorithm:

$$\mathbf{w}(1) = \mathbf{X}^{\#}\mathbf{b}(1)$$

vector **b**(1) – arbitrary but such that $b_i > 0$
i = 1, 2, ..., n

$$\mathbf{w}(k+1) = \mathbf{w}(k) + c\mathbf{X}^{\#}[\mathbf{e}(k) + |\mathbf{e}(k)|]$$

$$\mathbf{e}(k) = \mathbf{X}\mathbf{w}(k) - \mathbf{b}(k)$$

- when the inequalities $\mathbf{X}\mathbf{w} > \mathbf{0}$ have solution the algorithm converges for $0 < c \leq 1$

- if *all* the components of **e**(k) cease to be positive (*but not all zero*) at any iteration step, this indicates that the classes are not linear separable

- When $\mathbf{e}(k) = \mathbf{0} \rightarrow$ the $\mathbf{w}(k)$ is a solution

$$\mathbf{e}(k) = \mathbf{X}\mathbf{w}(k) - \mathbf{b}(k)$$

$$\mathbf{X}\,\mathbf{w} = \mathbf{b}$$

Example:

-Training set: $\omega_1 = \{(0, 0)^T, (0, 1)^T\}$

$\omega_2 = \{(1, 0)^T, (1, 1)^T\}$

Augmented pattern vectors:

$\omega_1 = \{(0, 0, 1)^T, (0, 1, 1)^T\}$

$\omega_2 = \{(1, 0, 1)^T, (1, 1, 1)^T\}$

Pattern vectors from ω_2 have to be multiplied by (-1)

$\omega_2 = \{(-1, 0, -1)^T, (-1, -1, -1)^T\}$

$\mathbf{b}(1) = (1, 1, 1, 1)^T$ and $c = 1$

Form the matrix $\mathbf{X}$:

$$\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

- Generalized inverse $X^{\#} = (\boldsymbol{X}^{\boldsymbol{T}}\mathbf{X})^{-\mathbf{1}}\,\mathbf{X}^{\boldsymbol{T}}$

$$\boldsymbol{X}^{\#} = \frac{1}{2}\begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ 3/2 & 1/2 & -1/2 & 1/2 \end{bmatrix}$$

1. Step

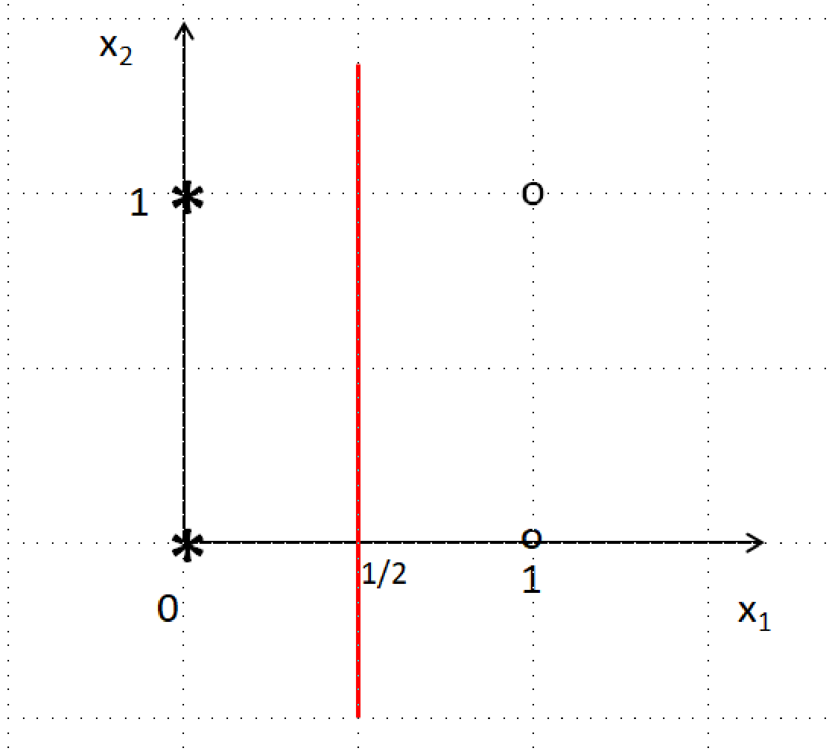
$$\mathbf{w}(1) = \mathbf{X}^{\#}\mathbf{b}(1) = \frac{1}{2}\begin{bmatrix} -1 & -1 & -1 & -1 \\ -1 & 1 & 1 & -1 \\ 3/2 & 1/2 & -1/2 & 1/2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

$$\mathbf{e}(1) = \mathbf{X}\mathbf{w}(1) - \mathbf{b}(1) = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ -1 & 0 & -1 \\ -1 & -1 & -1 \end{bmatrix}\begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

$\mathbf{e}(1) = \mathbf{0}$ $\mathbf{w}(1)$ is solution!

$\mathbf{w}(1) = (-2, 0, 1)^T$

$d(\mathbf{x}) = -2x_1 + 1$



Example:

-Training set: $\omega_1 = \{(0, 0)^T, (1, 1)^T\}$

$\omega_2 = \{(0, 1)^T, (1, 0)^T\}$

Augmented pattern vectors:

$\omega_1 = \{(0, 0, 1)^T, (1, 1, 1)^T\}$

$\omega_2 = \{(0, 1, 1)^T, (1, 0, 1)^T\}$

Pattern vectors from ω_2 have to be multiplied by (-1)

$\omega_2 = \{(0, -1, -1)^T, (-1, 0, -1)^T\}$

$\mathbf{b}(1) = (1, 1, 1, 1)^T$ and $c = 1$

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & -1 & -1 \\ -1 & 0 & -1 \end{bmatrix}$$

Generalized inverse $X^{\#} = (X^T \mathbf{X})^{-1} \mathbf{X}^T$

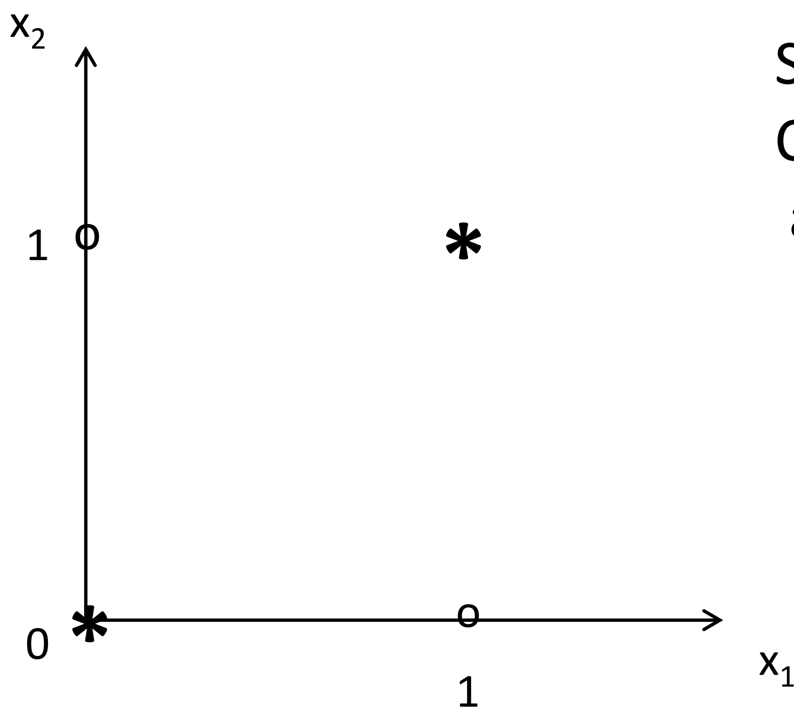
$$\mathbf{X}^{\#} = \frac{1}{2}\begin{bmatrix} -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 3/2 & -1/2 & -1/2 & -1/2 \end{bmatrix}$$

1. Step

$$\mathbf{w}(1) = \mathbf{X}^{\#}\mathbf{b}(1) = \frac{1}{2}\begin{bmatrix} -1 & 1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 3/2 & -1/2 & -1/2 & -1/2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\mathbf{e}(1) = \mathbf{X}\mathbf{w}(1) - \mathbf{b}(1) = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 0 & -1 & -1 \\ -1 & 0 & -1 \end{bmatrix}\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

$\mathbf{e}(1) = (-1, -1, -1, -1)^T$ – negative vector indicates that **Xw > 0** has no solution



$*$ $\in \omega_1$

o $\in \omega_2$

So-called XOR problem
Classes ω_1 and ω_2
are not linearly separable