Support Vector Machines

Outline

1. Separating hyperplane
2. Optimal hyperplane for linearly separable patterns
3. Optimal hyperplane for linearly nonseparable patterns
4. Building SVM (support vector machine) for pattern recognition
5. SVM for nonlinear regression
6. SVM for handling imbalanced data

1. Separating hyperplane

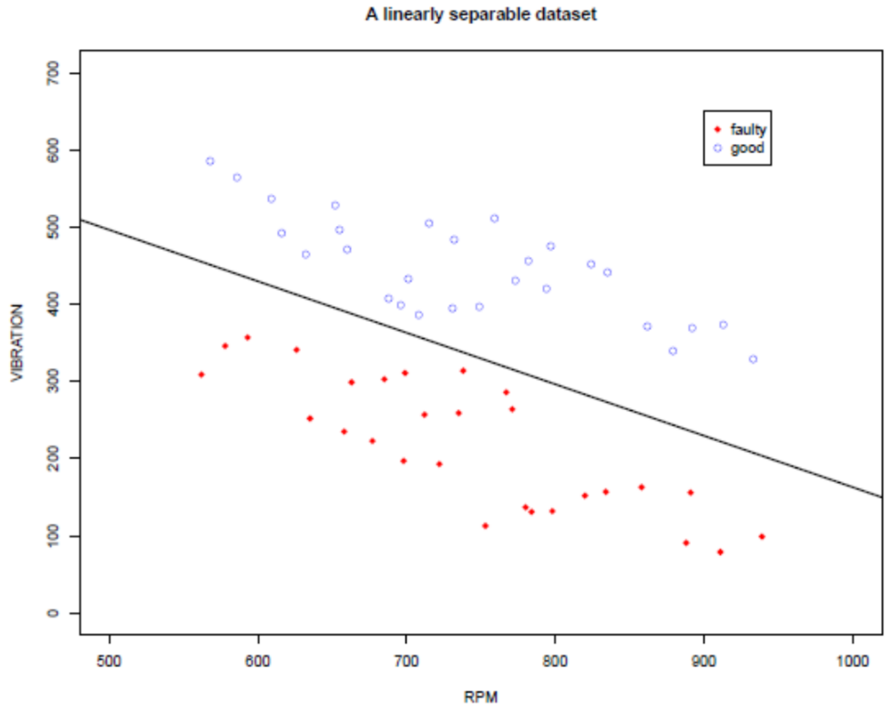
Linearly separable patterns can be separated by a hyperplane: Generator dataset.

ID	RPM	Vibration	Status
1	568	585	Good
2	586	565	Good
3	609	536	Good
….	…..	…..	…..
26	892	370	Good
27	913	373	Good
28	933	330	Good

ID	RPM	Vibration	Status
29	562	309	Faulty
30	578	346	Faulty
31	593	357	Faulty
….	…..	…..	…..
54	891	156	Faulty
55	911	79	Faulty
56	939	99	Faulty

Scatter plot and <u>regression line</u>:

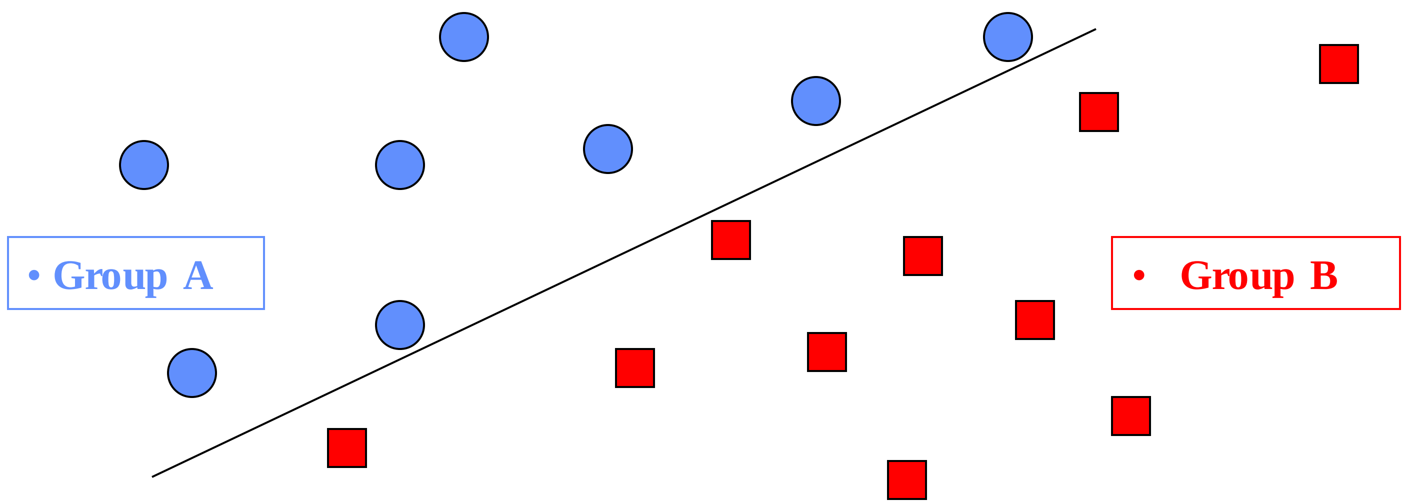
$$\text{Vibration} = 830 - 0.667 \times \text{RPM}$$



1. Separating hyperplane

Linear programming for classification

Consider 2 groups of data samples, group A and group B. They are linearly separable if there exists a hyperplane that separates these two groups such that all group A samples are on one side of the plane, and all group B samples are on the other side of the plane.



$$w_1x_1 + w_2x_2 + w_3x_3 + \ldots\ldots\ w_nx_n = c$$

1. Separating hyperplane

Linear programming for classification

LP can be used to minimize the <u>maximum deviation</u>:

minimize a

subject to

$$w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n \geq c - a, \qquad \text{for} \quad 1 \leq i \leq n_{gA}$$

$$w_1x_1 + w_2x_2 + \ldots\ldots + w_nx_n \leq c + a, \qquad \text{for } n_{gA} + 1 \leq i \leq n_{gA} + n_{gB}$$

$$a \geq 0$$

What is the role of "deviation"?

- n_{gA} = number of samples in Group A
- n_{gB} = number of samples in Group B
- a is the <u>deviation</u>
- c is a parameter, use fixed cut-off c but experiment with both negative and positive c

1. Separating hyperplane

Linearly separable patterns can be separated by a hyperplane: Generator dataset

ID	RPM	Vibration	Status
1	568	585	Good
2	586	565	Good
3	609	536	Good
….	…..	…..	…..
26	892	370	Good
27	913	373	Good
28	933	330	Good

ID	RPM	Vibration	Status
29	562	309	Faulty
30	578	346	Faulty
31	593	357	Faulty
….	…..	…..	…..
54	891	156	Faulty
55	911	79	Faulty
56	939	99	Faulty

minimize a

subject to

$$568\,w_1 + 585\,w_2 \geq c - a$$
$$586\,w_1 + 565\,w_2 \geq c - a$$
$$\ldots$$
$$933\,w_1 + 339\,w_2 \geq c - a$$

Group A

$$562\,w_1 + 309\,w_2 \leq c + a$$
$$578\,w_1 + 346\,w_2 \leq c + a$$
$$\ldots..$$
$$939\,w_1 + 99\,w_2 \leq c + a$$

Group B

$a \geq 0$ **deviation**

1. Separating hyperplane

LP solution for the linearly separable Generator dataset

minimize a

subject to

$$568\,w_1 + 585\,w_2 \geq c - a$$

$$586\,w_1 + 565\,w_2 \geq c - a$$

…

$$933\,w_1 + 339\,w_2 \geq c - a$$

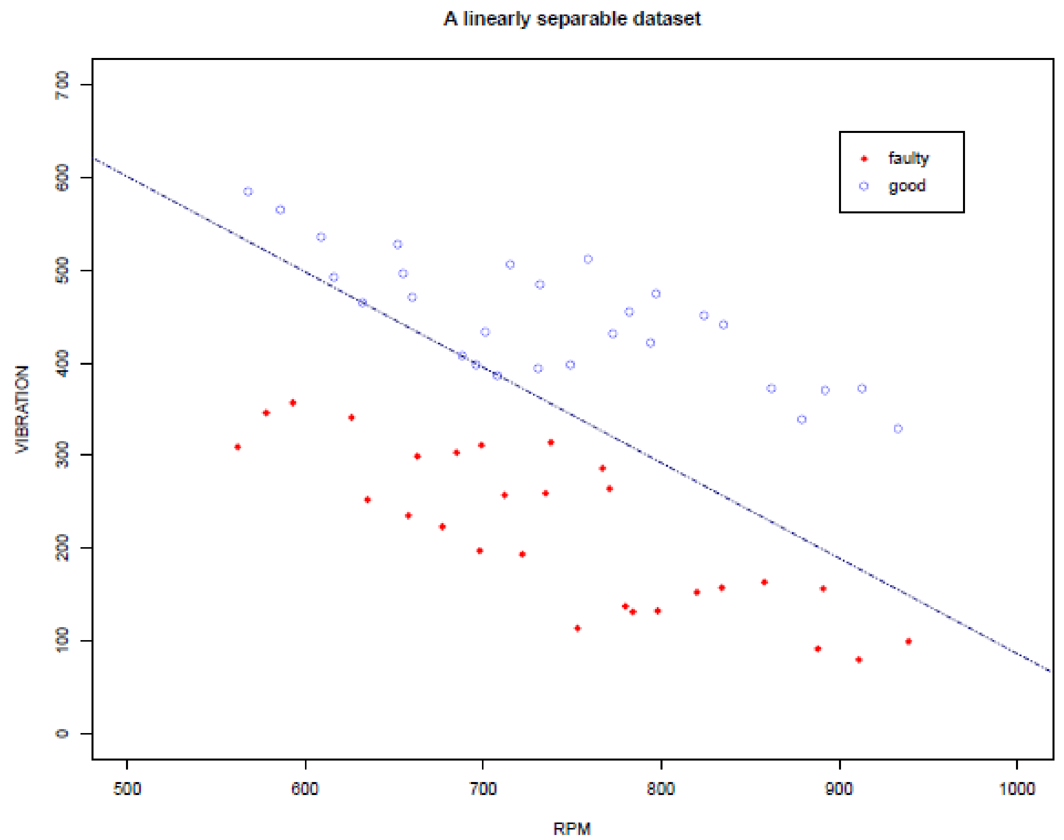
$$562\,w_1 + 309\,w_2 \leq c + a$$

$$578\,w_1 + 346\,w_2 \leq c + a$$

…..

$$939\,w_1 + 99\,w_2 \leq c + a$$

$$a \geq 0$$



- With c = 1000, the solution is $w_1 = 0.923439$, $w_2 = 0.895456$, **a = 0** ⇦ linearly separable data
- Two ‘good’ data points determine the solution: (632,465) and (696,399)

1. Separating hyperplane

Finding a separating hyperplane by solving a Quadratic programming problem

minimize $\frac{1}{2}(w_1^2 + w_2^2) = \frac{1}{2}\|w\|^2$

subject to

$w_0 + 568\, w_1 + 585\, w_2 \geq +1$

$w_0 + 586\, w_1 + 565\, w_2 \geq +1$

…

$w_0 + 933\, w_1 + 339\, w_2 \geq +1$

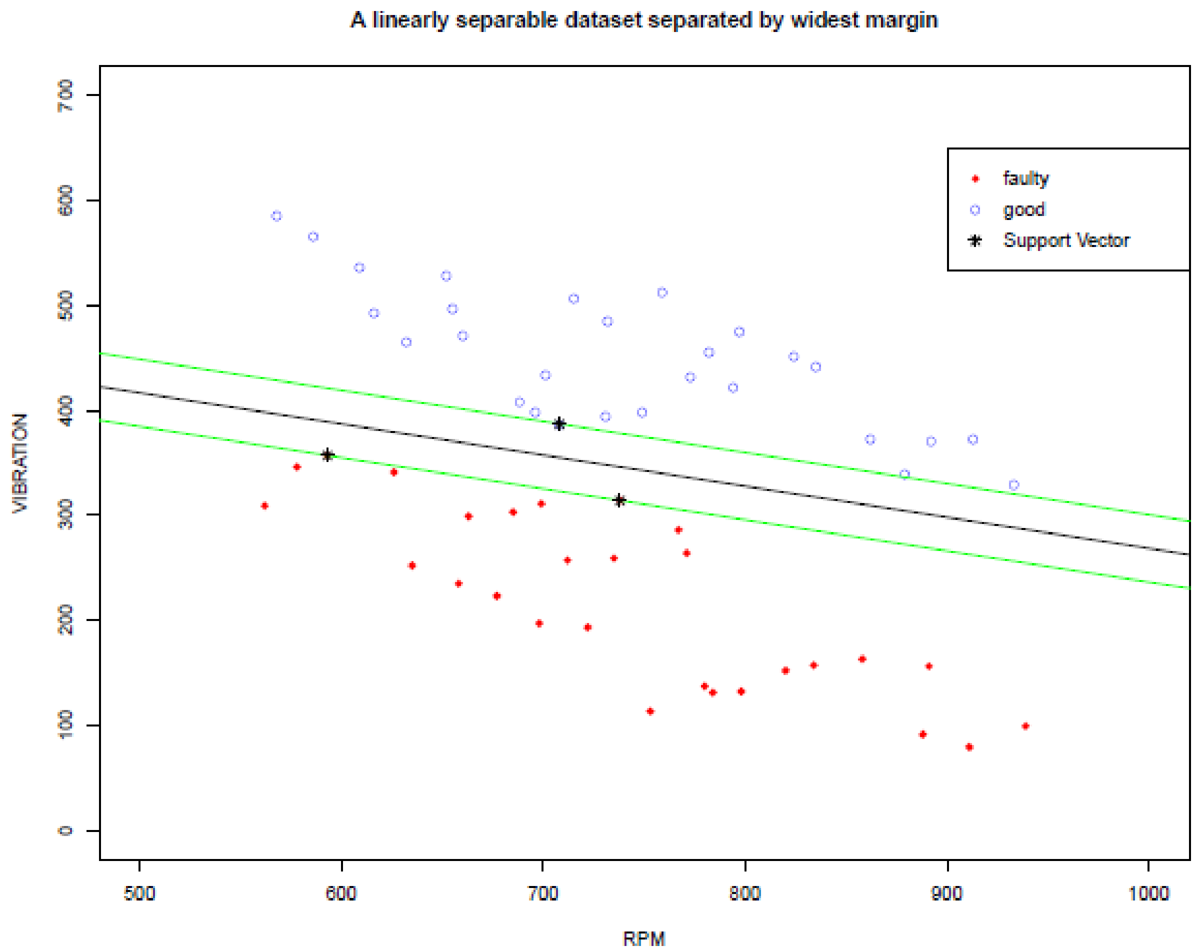
$w_0 + 562\, w_1 + 309\, w_2 \leq -1$

$w_0 + 578\, w_1 + 346\, w_2 \leq -1$

…..

$w_0 + 939\, w_1 + 99\, w_2 \leq -1$

This approach is Support Vector Machines (SVM)



1. Separating hyperplane

Finding a separating hyperplane by solving a Quadratic programming problem

Solution:

w_0 = -17.6249

w_1 = 0.00925

w_2 = 0.03120

Three data points are found to be support vectors:

(◆ 593,357):

$w_0 + 593 \times w_1 + 357 \times w_2 = -1$

(◆ 738,314):

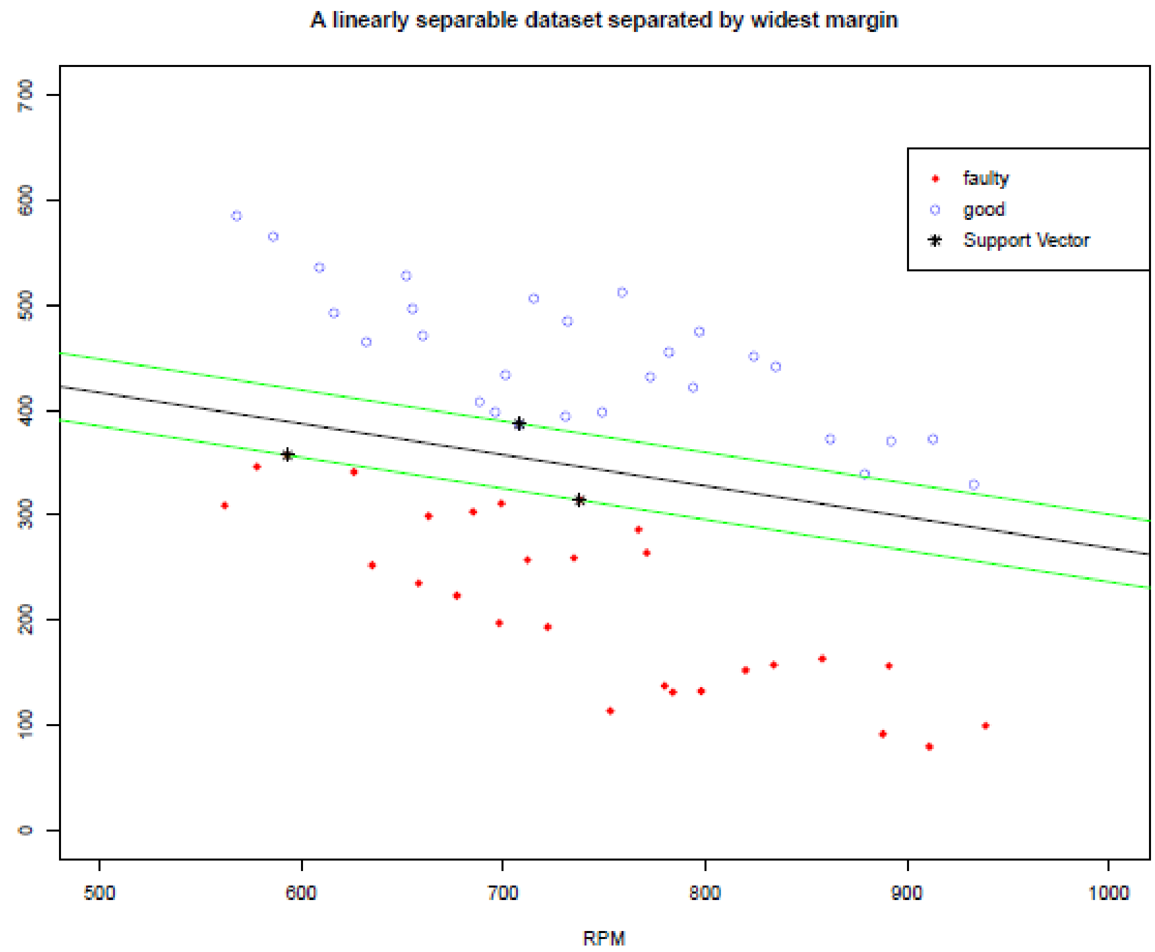
$w_0 + 738 \times w_1 + 314 \times w_2 = -1$

(○ 708,387):

$w_0 + 708 \times w_1 + 387 \times w_2 = +1$

Decision boundary:

$w_0 + w_1 \times \text{RPM} + w_2 \times \text{Vibration} = 0$



2. Optimal hyperplane for linearly separable patterns

- Let the training samples be $\mathbf{x_i}$, $i = 1,2,\ldots, N$, where x_i is n-dimensional.

- Associated with each $\mathbf{x_i}$ is a target value d_i with value of -1 or 1.

- The class represented by the subset with $d_i = -1$ and the class represented by the subset with $d_i = +1$ are **linearly separable** if there exists ($\mathbf{w}$,b) such that

$$\mathbf{w}^T \mathbf{x_i} + b \geq 0 \quad \text{for } d_i = +1$$

$$\mathbf{w}^T \mathbf{x_i} + b < 0 \quad \text{for } d_i = -1$$

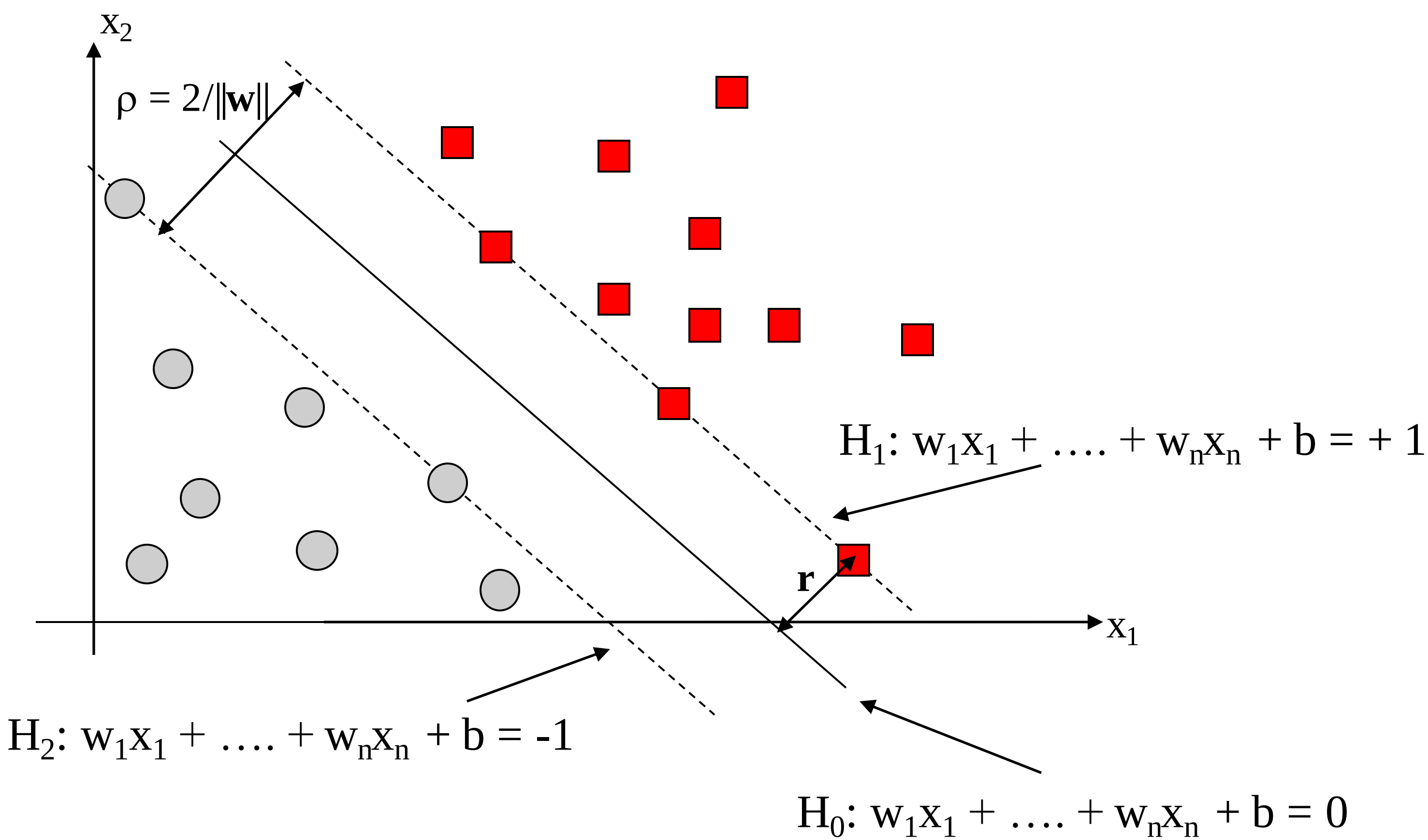
$$\boxed{b = w_0}$$

- The **margin of separation r** is the separation between the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ and the closest data point.

- The goal of a support vector machine is to find the **optimal hyperplane with the maximum margin of separation**.

2. Optimal hyperplane for linearly separable patterns

Consider the hyperplane that maximizes the margin of separation



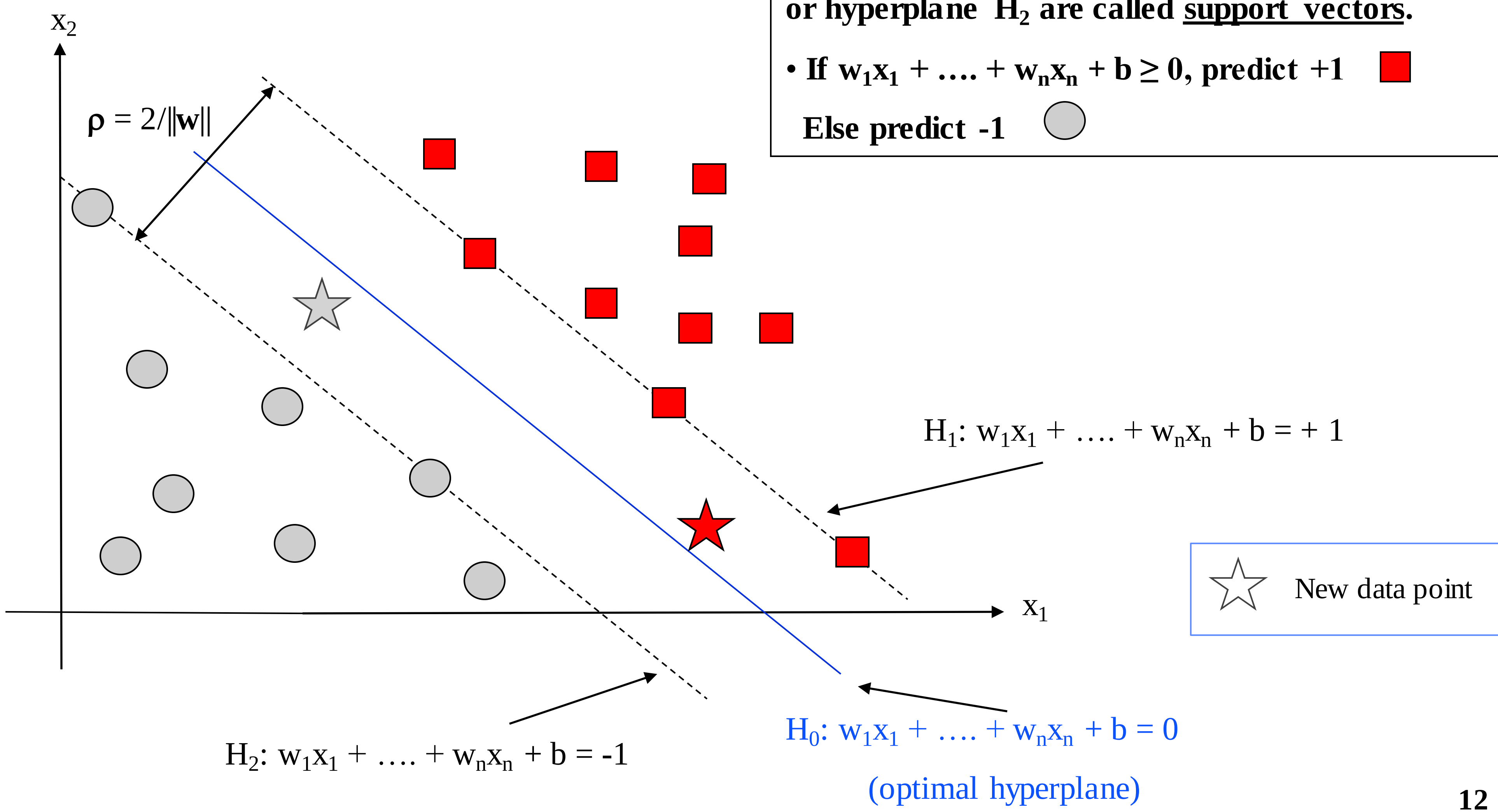
$\boxed{n = 2}$

2. Optimal hyperplane for linearly separable

Support Vector Machine (SVM):

- **Training data samples that lie on hyperplane H_1 or hyperplane H_2 are called <u>support vectors</u>.**
- **If $w_1x_1 + \ldots + w_nx_n + b \geq 0$, predict +1**

  Else predict -1



2. Optimal hyperplane for linearly separable patterns

- How to compute the margin of separation?
- Define discriminant function

  $g(\mathbf{x}) = \mathbf{w}_*^T\mathbf{x} + b_*$ and let

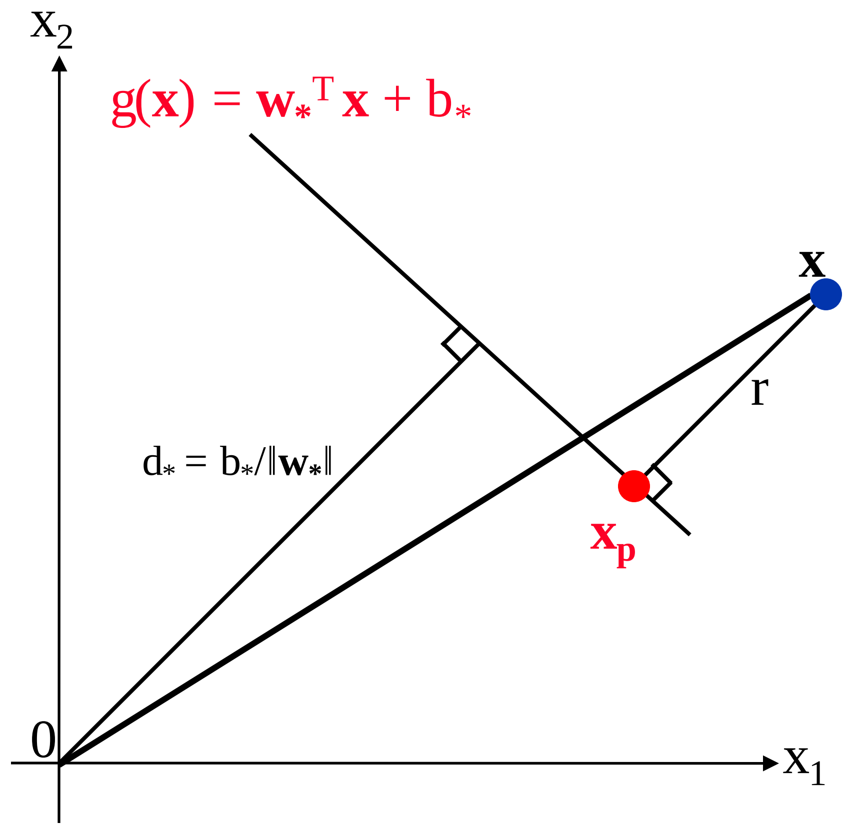
  $\mathbf{x} = \mathbf{x_p} + r\, \mathbf{w}_*/\|\mathbf{w}_*\|$

  where $\mathbf{x_p}$ is the normal projection of $\mathbf{x}$ onto the optimal hyperplane.

- r is the distance, r is positive if **x** is on the positive side of the optimal hyperplane, negative otherwise.
- Since $g(\mathbf{x_p}) = 0$, it follows that

  $g(\mathbf{x}) = \mathbf{w}_*^T\mathbf{x} + b_* = r\,\|\mathbf{w}_*\|$ or

  $r = g(\mathbf{x})/\|\mathbf{w}_*\|$



Details on next slide

2. Optimal hyperplane for linearly separable patterns

Note:

- $\mathbf{x} = \mathbf{x_p} + r\, \mathbf{w_*}/\|\mathbf{w_*}\|$
- $g(\mathbf{x}) = \mathbf{w_*}^T\mathbf{x} + b_*$
- $\mathbf{w_*}^T (\mathbf{x} - \mathbf{x_p}) = r\, \mathbf{w_*}^T \mathbf{w_*}/\|\mathbf{w_*}\|$
- $\mathbf{w_*}^T\mathbf{x} - \mathbf{w_*}^T\mathbf{x_p} = r\ \|\mathbf{w_*}\|$
- $\mathbf{w_*}^T\mathbf{x} + b_* = r\ \|\mathbf{w_*}\|$

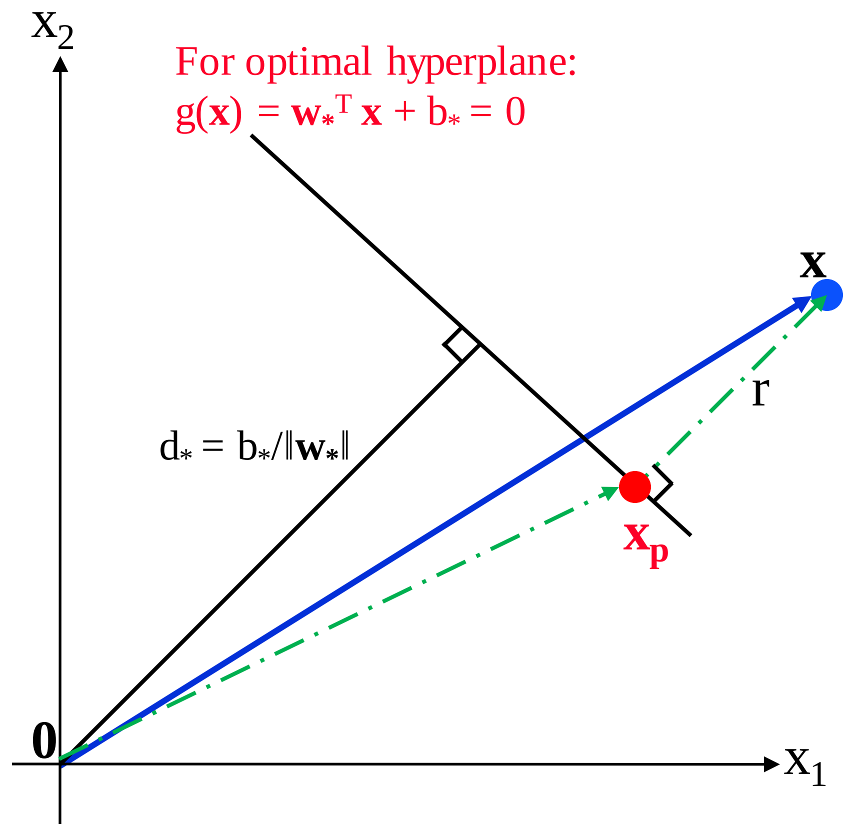
(since $g(\mathbf{x_p}) = 0$ by definition, $0 = \mathbf{w_*}^T\mathbf{x_p} + b_*$)

- Hence,

$$r = (\mathbf{w_*}^T\mathbf{x} + b_*)/\ \|\mathbf{w_*}\| = g(\mathbf{x})/\ \|\mathbf{w_*}\|$$

- Setting $\mathbf{x} = \mathbf{0}$ for the origin, we have the distance

from the origin to the optimal hyperplane $d_* = b_*/\ \|\mathbf{w_*}\|$



Note (on the note):

- Suppose **w** is an n-dimensional vector

$$\mathbf{w} = (w_1, w_2, \ldots, w_n)^T$$

- Then $\|\mathbf{w}\| = \text{sqrt}(w_1^2 + w_2^2 + \ldots + w_n^2)$

and $\|\mathbf{w}\|^2 = (w_1^2 + w_2^2 + \ldots + w_n^2) = \mathbf{w}^T\mathbf{w}$

- If $\mathbf{v} = \mathbf{w}/\|\mathbf{w}\|$

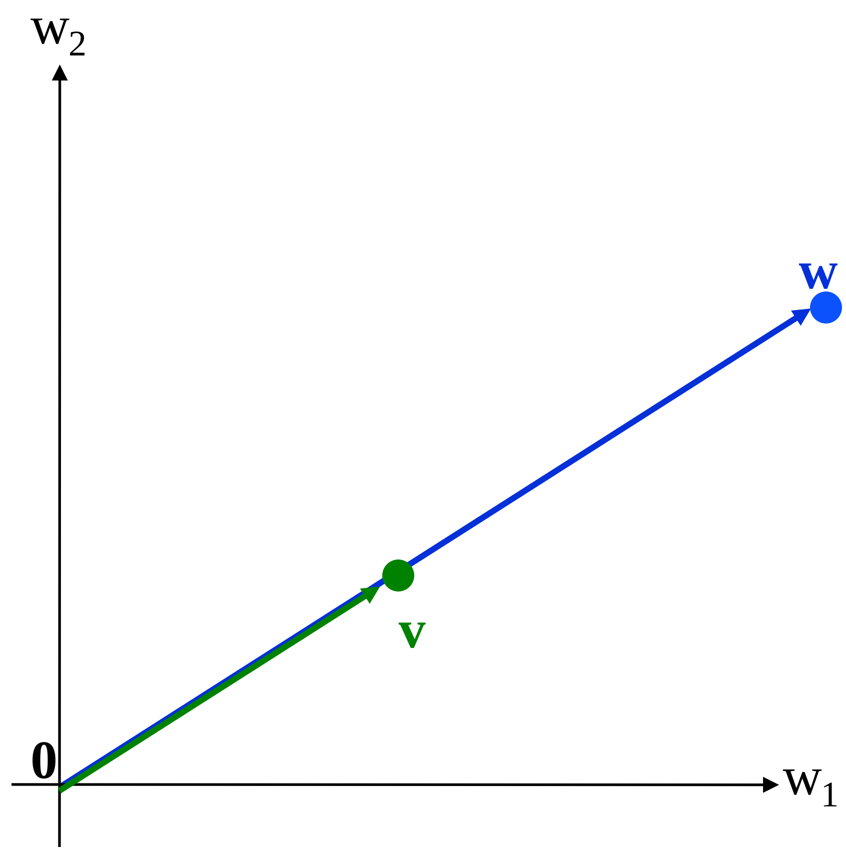
then $\|\mathbf{v}\| = \text{sqrt}(v_1^2 + v_2^2 + \ldots + v_n^2)$

$$= \text{sqrt}[(w_1^2 + w_2^2 + \ldots + w_n^2)/\|\mathbf{w}\|^2]$$

$$= \text{sqrt}(w_1^2 + w_2^2 + \ldots + w_n^2)/\|\mathbf{w}\|$$

$$= 1$$

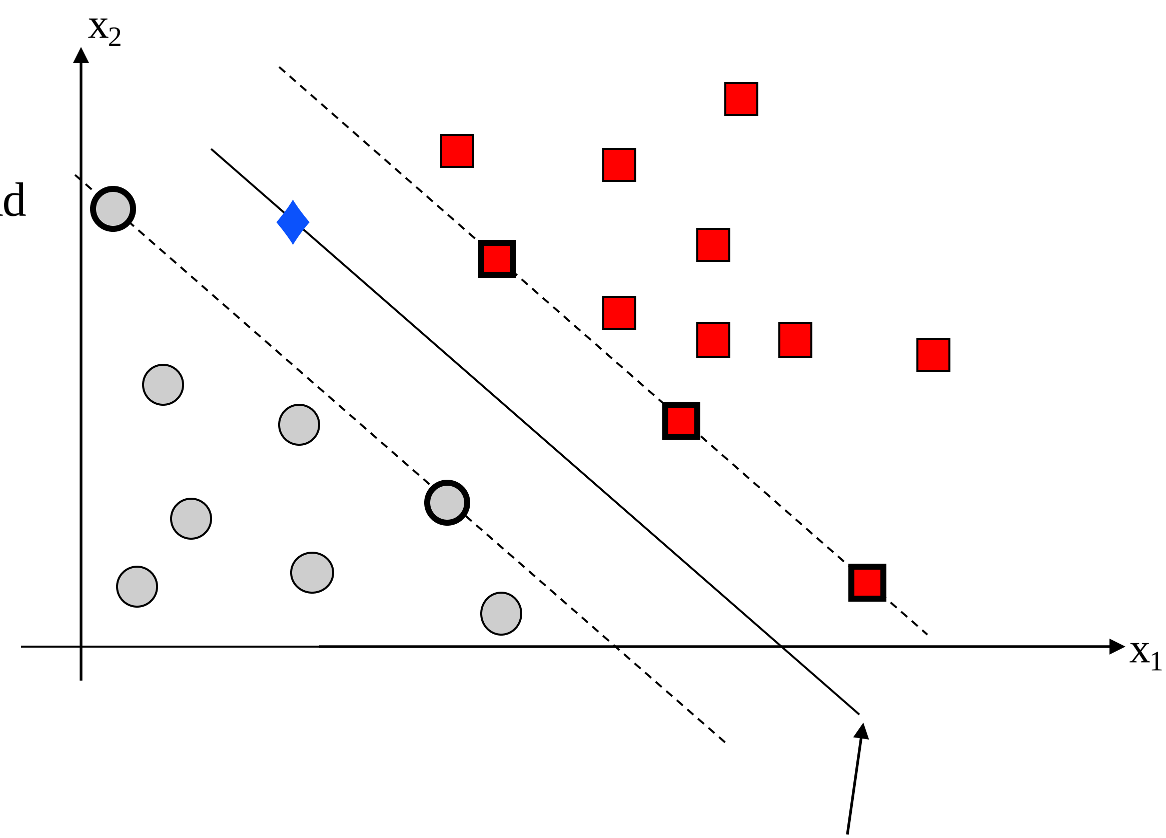
that is, **v** is a unit vector.



Another note (on the note):

- $\mathbf{w}_*$ is the <u>optimal</u> weight vector

  and b_* is its corresponding bias/threshold
- Given the values of $\mathbf{x} = (x_1, x_2 \ldots, x_n)$:
  - for $g(\mathbf{x}) = +1$
  - for $g(\mathbf{x}) > +1$
  - for $g(\mathbf{x}) = -1$
  - for $g(\mathbf{x}) < -1$
  - for $g(\mathbf{x}) = 0$



H_0: $w_1 x_1 + \ldots. + w_n x_n + b = 0$

$$\boxed{\begin{aligned} g(\mathbf{x}) &= \mathbf{w}_*^T \mathbf{x} + b_* \\ &= w_{*,1} x_1 + \ldots. + w_{*,n} x_n + b_* = 0 \end{aligned}}$$

2. Optimal hyperplane for linearly separable patterns

- The algebraic distance from the support vector $\mathbf{x}^{(s)}$ to the optimal hyperplane is

$$r = g(\mathbf{x}^{(s)}) / \|\mathbf{w}_*\| = 1/\|\mathbf{w}_0\| \text{ if } d^{(s)} = +1 \quad (\mathbf{w}_*^T \mathbf{x}^{(s)} + b_* = 1) \text{ and}$$

$$r = g(\mathbf{x}^{(s)}) / \|\mathbf{w}_*\| = -1/\|\mathbf{w}_0\| \text{ if } d^{(s)} = -1 \quad (\mathbf{w}_*^T \mathbf{x}^{(s)} + b_* = -1)$$

where the plus sign indicates $\mathbf{x}^{(s)}$ lies on the positive side of the hyperplane, and the minus sign indicates $\mathbf{x}^{(s)}$ lies on the negative side.

- Let ρ denote the optimum value of the margin of separation between the two classes of patterns, then

$$\rho = 2\,r = 2/\|\mathbf{w}_*\|$$

- Maximising the margin of separation is equivalent to minimizing the Euclidean norm of the weight vector $\mathbf{w}$.

2. Optimal hyperplane for linearly separable patterns

- Given the training sample $\{(\mathbf{x_i},d_i)\}$, i = 1, …., N, the **quadratic programming (QP)** problem is:

  minimise $\Phi(w) = ½\ \mathbf{w}^T\mathbf{w}$

  subject to $d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$ for i = 1, …., N

- The characteristics of the above QP are:

  convex quadratic objective function and linear constraints in **w**.

- The Lagrangian of the QP is

$$J(\mathbf{w},b,\boldsymbol{\alpha}) = ½\ \mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i[d_i(\mathbf{w}^T\mathbf{x_i} + b) - 1]$$

- $\alpha_i \geq 0$, i = 1, …, N is the Lagrange multiplier for constraint i.

- Optimality conditions:

$$\partial J(\mathbf{w},b,\alpha) / \partial \mathbf{w} = 0 \qquad \mathbf{w} = \sum_{i=1}^{N} \alpha_i\, d_i\, \mathbf{x_i}$$

and

$$\partial J(\mathbf{w},b,\boldsymbol{\alpha}) / \partial b = 0 \qquad \sum_{i=1}^{N} \alpha_i\, d_i = 0$$

2. Optimal hyperplane for linearly separable patterns

- The two optimality conditions yield the following:

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \, d_i \, \mathbf{x_i} \qquad \text{and}$$

$$\sum_{i=1}^{N} \alpha_i \, d_i = 0$$

- Note that the solution vector **w** is defined in terms of a sum that involves the N training samples.

- The Kuhn-Tucker optimality conditions also require that

$$\alpha_i \, [d_i(\mathbf{w}^T\mathbf{x_i} + b) - 1] = 0 \text{ for } i = 1, \ldots, N.$$

this is called the complementarity condition.

- If $\alpha_i > 0$, then $d_i(\mathbf{w}^T\mathbf{x_i} + b) = 1$.
- If $d_i(\mathbf{w}^T\mathbf{x_i} + b) \neq 1$, then $\alpha_i = 0$

- Instead of solving QP, we could also solve its dual problem and obtain the same optimal value.

2. Optimal hyperplane for linearly separable patterns

The dual of QP can be obtained as follows:

- $J(w,b,\alpha) = \frac{1}{2}\ \mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N}\alpha_i[d_i(\mathbf{w}^T\mathbf{x_i} + b) - 1]$

$$= \underline{\tfrac{1}{2}\,\mathbf{w}^T\mathbf{w}} - \underline{\sum_{i=1}^{N}\alpha_i\, d_i\, \mathbf{w}^T\mathbf{x_i}} - b\underline{\sum_{i=1}^{N}\alpha_i\, d_i} + \underline{\sum_{i=1}^{N}\alpha_i}$$

but $\underline{\mathbf{w}^T\mathbf{w}} = \underline{\sum_{i=1}^{N}\alpha_i\, d_i\, \mathbf{w}^T\mathbf{x_i}} = \underline{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, d_i\, d_j\mathbf{x_i}^T\mathbf{x_j}}$

and $\underline{\sum_{i=1}^{N}\alpha_i\, d_i} = 0$ (the two optimality conditions on previous slide)

- Hence, we <u>maximise</u> the dual quadratic programming

$$Q(\boldsymbol{\alpha}) = \underline{\sum_{i=1}^{N}\alpha_i} - \tfrac{1}{2}\underline{\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\, d_i\, d_j\mathbf{x_i}^T\mathbf{x_j}}$$

subject to: $\sum_{i=1}^{N}\alpha_i\, d_i = 0$ and $\alpha_i \geq 0$ for $i = 1,2, \ldots, N$.

2. Optimal hyperplane for linearly separable patterns

- Note that the dual problem is given in terms of $\mathbf{x_i}$ and that the unknown variables are α_i , i = 1, …, N.

- Once the <u>optimal value variables</u> of dual QP, $\alpha_{*,i}$ have been computed, the optimum weight vector $\mathbf{w_o}$ is equal to

$$\mathbf{w}_* = \sum_{i=1}^{N} \alpha_{*,i} d_i \; \mathbf{x_i}$$

and the optimum bias b_* is

$$b_* = 1 - \mathbf{w}_*^T \, \mathbf{x^{(s)}} \text{ for } d^{(s)} = +1 \qquad \text{where } \mathbf{x^{(s)}} \text{ is a support vector}$$

<u>Notation used</u>:

- for n-dimensional vectors **x** and **y**, $\mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} (x_i \times y_i)$
- $||\mathbf{x}||$ = the Euclidean distance of **x** from the origin

$$= \text{sqrt}[\sum_{i=1}^{n} x_i^2 \,] = \quad \text{sqrt}\,(x_1^2 + x_2^2 + \ldots.\; x_n^2)$$

2. Optimal hyperplane for linearly separable patterns

- Example.

i	x_1	x_2	Class	d_i
1	0	2	Bad	-1
2	2	2.4	Bad	-1
3	2	5	Good	+1
4	3	4	Good	+1
5	1	4.5	Good	+1
6	1	4	Good	+1
7	1	2	Bad	-1
8	0	1.2	Bad	-1

minimize $\frac{1}{2}(w_1^2 + w_2^2) = \frac{1}{2}\|w\|^2$

subject to

i=1: $0\,w_1 + 2\,w_2 + b \le -1$

i=2: $2\,w_1 + 2.4\,w_2 + b \le -1$

i=7: $1\,w_1 + 2\,w_2 + b \le -1$

i=8: $0\,w_1 + 1.2\,w_2 + b \le -1$

i=3: $2\,w_1 + 5\,w_2 + b \ge +1$

i=4: $3\,w_1 + 4\,w_2 + b \ge +1$

i=5: $1\,w_1 + 4.5\,w_2 + b \ge +1$

i=6: $1\,w_1 + 4\,w_2 + b \ge +1$

2. Optimal hyperplane for linearly separable patterns

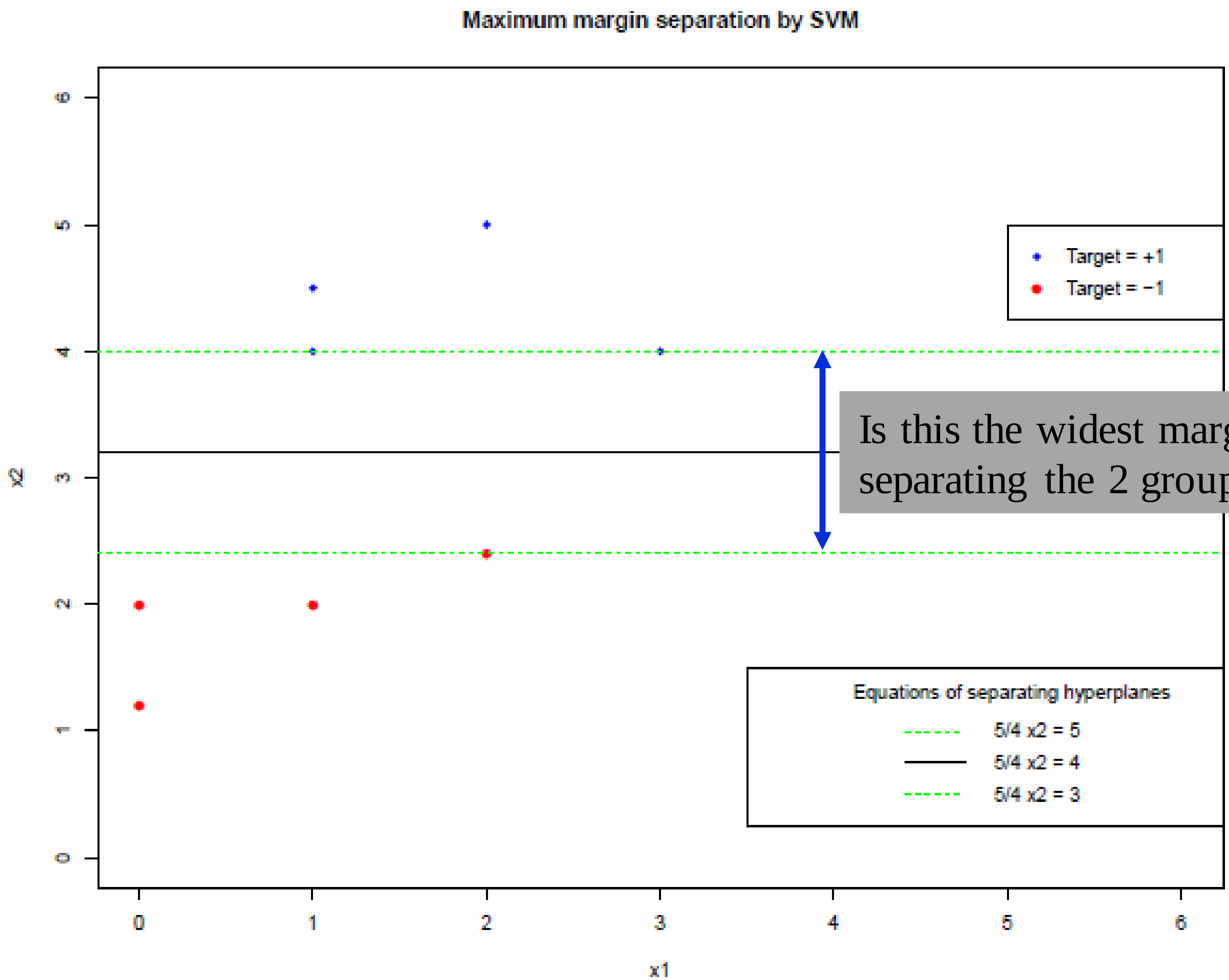
- Example.

QP: minimize $\frac{1}{2}(w_1^2 + w_2^2) = \frac{1}{2}\|w\|^2$

subject to

$$\begin{aligned}
0\,w_1 + 2\,w_2 + b &\le -1\\
2\,w_1 + 2.4\,w_2 + b &\le -1\\
1\,w_1 + 2\,w_2 + b &\le -1\\
0\,w_1 + 1.2\,w_2 + b &\le -1
\end{aligned}$$

$$\begin{aligned}
2\,w_1 + 5\,w_2 + b &\ge +1\\
3\,w_1 + 4\,w_2 + b &\ge +1\\
1\,w_1 + 4.5\,w_2 + b &\ge +1\\
1\,w_1 + 4\,w_2 + b &\ge +1
\end{aligned}$$



Is this the widest margin separating the 2 groups?

Check if $w_1 = 0$, $w_2 = 5/4$ and $b = -4$ satisfy all the optimality conditions (page 22)

2. Optimal hyperplane for linearly separable patterns

minimize $\frac{1}{2}(w_1^2 + w_2^2) = \frac{1}{2}\|w\|^2$

subject to

$$\begin{aligned} 0\,w_1 + 2\,w_2 + b &\le -1 \\ 2\,w_1 + 2.4\,w_2 + b &\le -1 \\ 1\,w_1 + 2\,w_2 + b &\le -1 \\ 0\,w_1 + 1.2\,w_2 + b &\le -1 \end{aligned}$$

$$\begin{aligned} 2\,w_1 + 5\,w_2 + b &\ge +1 \\ 3\,w_1 + 4\,w_2 + b &\ge +1 \\ 1\,w_1 + 4.5\,w_2 + b &\ge +1 \\ 1\,w_1 + 4\,w_2 + b &\ge +1 \end{aligned}$$

Check if $w_1 = 0$, $w_2 = 5/4$ and $b = -4$ satisfy all the optimality conditions (page 22)

- First check feasibility:

$$\begin{aligned} 0\,w_1 + 2\,w_2 + b &= 0 + 2(5/4) - 4 < -1 \\ 2\,w_1 + 2.4\,w_2 + b &= 0 + 2.4(5/4) - 4 = -1 \\ 1\,w_1 + 2\,w_2 + b &= 0 + 2(5/4) - 4 < -1 \\ 0\,w_1 + 1.2\,w_2 + b &= 0 + 1.2(5/4) - 4 < -1 \end{aligned}$$

✓

$$\begin{aligned} 2\,w_1 + 5\,w_2 + b &= 0 + 5(5/4) - 4 > +1 \\ 3\,w_1 + 4\,w_2 + b &= 0 + 4(5/4) - 4 = +1 \\ 1\,w_1 + 4.5\,w_2 + b &= 0 + 4.5(5/4) - 4 > +1 \\ 1\,w_1 + 4\,w_2 + b &= 0 + 4(5/4) - 4 = +1 \end{aligned}$$

✓

2. Optimal hyperplane for linearly separable patterns

$0\ w_1 + 2\ w_2 + b = 0 + 2(5/4) - 4 < -1$

$2\ w_1 + 2.4\ w_2 + b = 0 + 2.4(5/4) - 4 = -1$

$1\ w_1 + 2\ w_2 + b = 0 + 2(5/4) - 4 < -1$

$0\ w_1 + 1.2\ w_2 + b = 0 + 1.2(5/4) - 4 < -1$

$2\ w_1 + 5\ w_2 + b = 0 + 5(5/4) - 4 > +1$

$3\ w_1 + 4\ w_2 + b = 0 + 4(5/4) - 4 = +1$

$1\ w_1 + 4.5\ w_2 + b = 0 + 4.5(5/4) - 4 > +1$

$1\ w_1 + 4\ w_2 + b = 0 + 4(5/4) - 4 = +1$

- Then check complementarity conditions:

$$\alpha_i [d_i(w^T x_i + b) - 1] = 0 \text{ for } i = 1, \ldots, N.$$

For samples with $d_i = -1$:

$0\ w_1 + 2\ w_2 + b < -1 \rightarrow$ let $\alpha_1 = 0$

$2\ w_1 + 2.4\ w_2 + b = -1 \rightarrow \alpha_2 = ?$

$1\ w_1 + 2\ w_2 + b < -1 \rightarrow$ let $\alpha_7 = 0$

$0\ w_1 + 1.2\ w_2 + b < -1 \rightarrow$ let $\alpha_8 = 0$

For samples with $d_i = +1$:

$2\ w_1 + 5\ w_2 + b > +1 \rightarrow$ let $\alpha_3 = 0$

$3\ w_1 + 4\ w_2 + b = +1 \rightarrow$ let $\alpha_4 = ?$

$1\ w_1 + 4.5\ w_2 + b > +1 \rightarrow$ let $\alpha_5 = 0$

$1\ w_1 + 4\ w_2 + b = +1 \rightarrow$ let $\alpha_6 = ?$

2. Optimal hyperplane for linearly separable patterns

For samples with $d_i = -1$:

$0\ w_1 + 2\ w_2 + b < -1 \rightarrow$ let $\alpha_1 = 0$

$2\ w_1 + 2.4\ w_2 + b = -1 \rightarrow \alpha_2 = ?$

$1\ w_1 + 2\ w_2 + b < -1 \rightarrow$ let $\alpha_7 = 0$

$0\ w_1 + 1.2\ w_2 + b < -1 \rightarrow$ let $\alpha_8 = 0$

For samples with $d_i = +1$:

$2\ w_1 + 5\ w_2 + b > +1 \rightarrow$ let $\alpha_3 = 0$

$3\ w_1 + 4\ w_2 + b = +1 \rightarrow$ let $\alpha_4 = ?$

$1\ w_1 + 4.5\ w_2 + b > +1 \rightarrow$ let $\alpha_5 = 0$

$1\ w_1 + 4\ w_2 + b = +1 \rightarrow$ let $\alpha_6 = ?$

- Are there values of $\alpha_2, \alpha_4, \alpha_6 \geq 0$ such that:

$$w = \sum_{i=1}^{8} \alpha_i d_i x_i \quad \text{and} \quad \sum_{i=1}^{8} \alpha_i d_i = 0$$

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 5/4 \end{pmatrix} = -\alpha_2 \begin{pmatrix} 2 \\ 2.4 \end{pmatrix} + \alpha_4 \begin{pmatrix} 3 \\ 4 \end{pmatrix} + \alpha_6 \begin{pmatrix} 1 \\ 4 \end{pmatrix}$$

$$-\alpha_2 + \alpha_4 + \alpha_6 = 0 \qquad ?$$

- Answer: $\alpha_2 = 50/64$, $\alpha_4 = 25/64$, $\alpha_6 = 25/64$

Conclusion:

- $w_1 = 0$, $w_2 = 5/4$ and $b = -4$ is the optimal solution to QP.
- $5/4\ x_2 - 4 = 0$ is the optimal separating hyperplane.
- Samples i = 2,4,6 are support vectors.
- $\|w\| = \text{sqrt}(0^2 + (5/4)^2) = 5/4$
- $g\left(\begin{pmatrix} 1 \\ 4 \end{pmatrix}\right) = 1$, distance to optimal hyperplane = $g(x)/\|w\| = 4/5$.

3. Optimal hyperplane for linearly nonseparable patterns

- We consider now the case of nonseparable patterns where it is not possible to construct a hyperplane without encountering classification errors.
- The margin of separation between classes is soft if a data point $(\mathbf{x_i}, d_i)$ violates the condition $d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1$ for $i = 1, \ldots, N$
- We introduce a set of nonnegative scalars ξ_i, $i = 1, \ldots, N$ and have the new condition

$$d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, N$$

- ξ_i is also called a slack variable.
- For $0 \leq \xi_i \leq 1$, the data point falls inside the region of separation, but on the right side of the decision surface (optimal hyperplane).
- For $\xi_i > 1$, it falls on the wrong side of the separating hyperplane.

3. Optimal hyperplane for linearly nonseparable patterns

- For this case, the support vectors are data points that satisfy

$$d_i(\mathbf{w}^T\mathbf{x_i} + b) = 1 - \xi_i, \; i = 1, \ldots, N$$

- ξ_i is a slack variable.

- The goal now is to find a separating hyperplane that minimises the classification errors, for example:

$$\text{minimise } \Phi(\xi) = \sum_{i=1}^{N} I(\xi_i - 1)$$

where $I(\xi)$ is an <u>indicator function</u> and $I(\xi) = 0$ if $\xi \leq 0$, and 1 otherwise.

- The minimisation problem becomes non-convex and it is NP-complete.

> **• NP = "Nondeterministic Polynomial"**
>
> **• NP complete problem: no polynomial time algorithms are known to solve this problem**

3. Optimal hyperplane for linearly nonseparable patterns

- We approximate the misclassification counts by

$$\Phi(\boldsymbol{\xi}) = \sum_{i=1}^{N} \xi_i$$

- The primal quadratic programming problem for nonseparable case is

$$\text{minimise } \Phi(w, \xi) = \tfrac{1}{2}\ \mathbf{w}^T\mathbf{w} + C \sum_{i=1}^{N} \xi_i$$

$$\text{subject to } d_i(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i$$

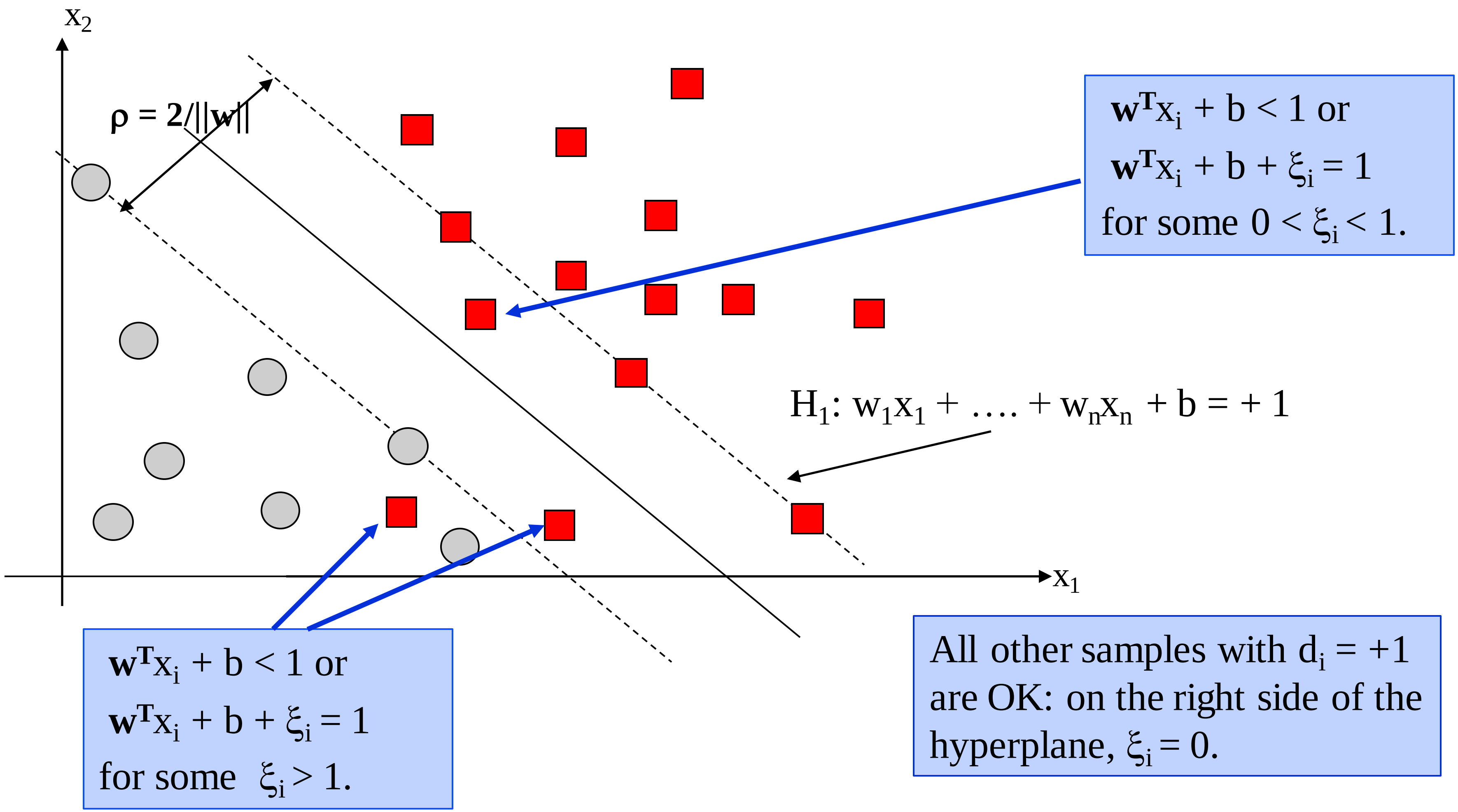
$$\text{and} \quad \xi_i \geq 0 \qquad\qquad \text{for } i = 1, \ldots., N$$

where C is a user-specified parameter that can be determined experimentally.

3. Optimal hyperplane for linearly nonseparable patterns

- Consider samples with $d_i = +1$

we want $(\mathbf{w}^T\mathbf{x_i} + b) \geq 1 - \xi_i$ with $\xi_i = 0$ if possible, if not $\xi_i > 0$



3. Optimal hyperplane for linearly nonseparable patterns

- The dual of the above QP is maximize

$$Q(\alpha) = \sum_{i=1}^{N} \alpha_i \; - ½ \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j \mathbf{x_i}^T \mathbf{x_j}$$

subject to the constraints:

$$\sum_{i=1}^{N} \alpha_i d_i = 0 \quad \text{and} \quad 0 \le \alpha_i \le C \quad \text{for } i = 1,2, \ldots, N.$$

- The optimum solution for the weight vector w is given by

$$\mathbf{w_*} = \sum_{i=1}^{Ns} \alpha_{*,i} d_i \mathbf{x_i}$$ where N_s is the number of support vectors.

3. Optimal hyperplane for linearly nonseparable patterns

- The complementarity conditions are

$$\alpha_i \, [d_i(\mathbf{w}^T\mathbf{x_i} + b) \; - 1 + \xi_i] = 0, \qquad i = 1,2, \ldots., N$$

$$\mu_i \, \xi_i \; = 0$$

$$\xi_i \, (C - \alpha_i \,) = 0$$

(μ_i is Lagrange multiplier associated with the constraint $\xi_i \geq 0$ in the primal problem)

$\boldsymbol{\alpha_i > 0}$

- To determine the optimal bias b_*, take the points with $0 < \alpha_{*,i} < C$. For these points, $\xi_i = 0$, and compute b from the first complementarity condition above:

$$d_i(\mathbf{w}^T\mathbf{x_i} + b) \; = 1$$

and take the mean value of b resulting from all such data points in the training sample.

4. Building SVM for pattern recognition

- The construction of svm for a pattern recognition task hinges on two important ideas:

  1. Nonlinear mapping of an input vector into a high dimensional feature space that is hidden from both the input and the output.

  2. Construction of an optimal hyperplane for separating the features discovered in Step 1.

- Cover's theorem on the separability of patterns:

  A complex pattern-classification problem cast in a high dimensional space nonlinearly is more likely to be linearly separable than in a low dimensional space.

- The separating hyperplane is to be defined as a linear function of vectors drawn from the feature space rather than in the original space. How is this hyperplane to be constructed?

Inner-product kernel

- Let $\mathbf{x}$ denote a vector drawn from the m_0 dimensional input space.
- Let $\varphi_j(\mathbf{x})$, $j = 1,\ldots, m_1$ denote a set of nonlinear transformation from the input space to the m_1 dimensional feature space.
- It is assumed that $\varphi_j(\mathbf{x})$ is defined a priori for all $j = 1,\ldots, m_1$.
- Let $\varphi_0(\mathbf{x}) = 1$ for all $\mathbf{x}$.
- Define a decision hyperplane in the feature space as follows

$$\sum_{j=0}^{m_1} w_j\, \varphi_j(\mathbf{x}) = 0$$

(note that w_0 denotes the bias of the hyperplane).

- Let $\boldsymbol{\Psi}(\mathbf{x}) = [\varphi_0(\mathbf{x}), \varphi_1(\mathbf{x}), \ldots, \varphi_{m1}(\mathbf{x})]^T$

then the decision surface can be written as

$$w^T \boldsymbol{\Psi}(\mathbf{x}) = 0$$

- Example:

  $b + w^T \mathbf{x} = 5 + 2x_1 - 7x_2 = 0$

  $b = 5,\ w_1 = 2,\ w_2 = -7$

- Or similarly

  $\mathbf{w}^T \mathbf{y} = 0$, where

  $\mathbf{w} = [5\ \ 2\ -7]^t$

  $\mathbf{y} = [1\ \ x_1\ \ x_2]^t$

4. Building SVM for pattern recognition

Inner-product kernel (continued)

- In the new space, we search for linear separability of the features, that is we want a weight vector **w** such that

$$\mathbf{w} = \sum_{i=1}^{N} \alpha_i \, d_i \, \Psi(\mathbf{x_i})$$

Compare to the original equation for w on page 19

where $\Psi(\mathbf{x_i})$ is the feature vector which corresponds to the input pattern $\mathbf{x_i}$.

- The decision surface computed in the feature space is

$$\sum_{i=1}^{N} \alpha_i \, d_i \, \Psi^T(\mathbf{x_i}) \, \Psi(\mathbf{x}) = 0$$

This is just $w^T x = 0$ in the original space

- The term $\Psi^T(\mathbf{x_i}) \, \Psi(\mathbf{x})$ represents the inner product of two vectors induced in the feature space by the input vector **x** and the i-th input vector $\mathbf{x_i}$.

4. Building SVM for pattern recognition

Inner-product kernel (continued)

- Define the inner-product kernel $K(\mathbf{x},\mathbf{x}_i)$ as follows:

$$K(\mathbf{x}, \mathbf{x}_i) = \Psi^T(\mathbf{x})\,\Psi(\mathbf{x}_i)$$

$$= \sum_{j=0}^{m_1} \varphi_j(\mathbf{x})\, \varphi_j(\mathbf{x}_i) \qquad \text{for } i = 1,2\ldots, N$$

- The inner-product kernel is a symmetric function:

$$K(\mathbf{x}, \mathbf{x}_i) = K(\mathbf{x}_i, \mathbf{x}) \quad \text{for all } i.$$

- Using the inner-product kernel $K(\mathbf{x}, \mathbf{x}_i)$, we may construct the optimal hyperplane in the feature space without having to consider the feature space itself in explicit form. The optimal hyperplane is

$$\sum_{i=1}^{N} \alpha_i\, d_i\, K(\mathbf{x}, \mathbf{x}_i) = 0$$

- **$\mathbf{x}_i$ is the i-th "training" data sample**
- **$\mathbf{x}$ is a new point with "unknown" target value.**

4. Building SVM for pattern recognition

Optimum design of SVM

- Given the training sample $\{(\mathbf{x_i}, d_i)\}$, i = 1,2,….,N, find the Lagrange multipliers $\{\alpha_i\}$, i = 1,2,…., N that maximise the objective function

- $$Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \tfrac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j d_i d_j K(\mathbf{x_i}^T, \mathbf{x_j})$$ $\boxed{\text{Compare with dual QP on pg 31}}$

  subject to the constraints:

  $$\sum_{i=1}^{N} \alpha_i d_i = 0 \text{ and } 0 \le \alpha_i \le C \quad \text{for } i = 1,2, \ldots, N$$

  where C is a user-specified positive parameter.

- We may view $K(\mathbf{x_i}^T, \mathbf{x_j})$ as the ij-th element of a symmetric N-by-N matrix as shown by $\mathbb{K} = \{K(\mathbf{x_i}^T, \mathbf{x_j})\}$ for i, j = 1,2,….,N

Inner product kernel for 3 common types of SVM:

- Type: polynomial learning machine

  Inner product kernel: $K(\mathbf{x}, \mathbf{x_i}) = (\mathbf{x}^T \mathbf{x_i} + 1)^p$

  Comment: p is specified a priori by the user.

- Type: Radial basis function network.

  Inner product kernel: $K(\mathbf{x}, \mathbf{x_i}) = \exp(-1/(\sigma^2) \| \mathbf{x} - \mathbf{x_i}\|^2)$

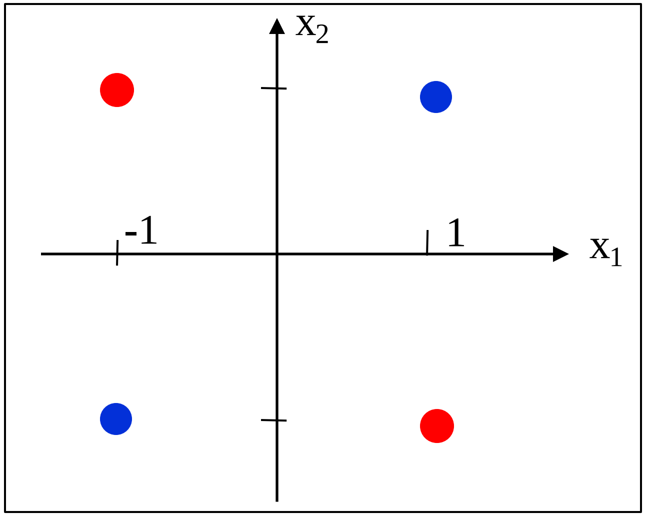
  Comment: the width σ^2 is specified a priori by the user.

- Type: two-layer perceptron

  Inner product kernel: $\tanh(\beta_0 \mathbf{x}^T \mathbf{x_i} + \beta_1)$

  Comment: only for some values of β_0 and β_1 .

Example: XOR problem



XOR problem:		
i	Input vector $\mathbf{x_i}$	Desired response d_i
1	(-1,-1)	-1
2	(-1,+1)	+1
3	(+1,-1)	+1
4	(+1,+1)	-1

- Let $K(\mathbf{x}, \mathbf{x_i}) = (\mathbf{x}^T\mathbf{x_i} + 1)^2$

$$= (x_1x_{i1} + x_2x_{i2} + 1)^2$$

$$= 1 + x_1^2x_{i1}^2 + 2x_1x_2x_{i1}x_{i2} + x_2^2x_{i2}^2 + 2x_1x_{i1} + 2x_2x_{i2}$$

2D samples mapped into 5D feature space

- Hence, $\Psi(\mathbf{x}) = [1, x_1^2, \sqrt{2}\, x_1x_2 , x_2^2 , \sqrt{2}\, x_1 , \sqrt{2}\, x_2]^T$

 and $\Psi(\mathbf{x_i}) = [1, x_{i1}^2, \sqrt{2}\, x_{i1}x_{i2} , x_{i2}^2 , \sqrt{2}\, x_{i1}, \sqrt{2}\, x_{i2}]^T$

- We also find $K(1,1) = [\,(-1,-1)^T(-1,-1) + 1)]^2 = 3^2 = 9.$

4. Building SVM for pattern recognition

Example: XOR problem

XOR problem:		
i	Input vector $\mathbf{x_i}$	Desired response $d_\mathbf{i}$
1	(-1,-1)	-1
2	(-1,+1)	+1
3	(+1,-1)	+1
4	(+1,+1)	-1

- $\Psi(\mathbf{x}) = [1, x_1^2, \sqrt{2}\, x_1 x_2, x_2^2, \sqrt{2}\, x_1, \sqrt{2}\, x_2]^T$

  and $\Psi(\mathbf{x_i}) = [1, x_{i1}^2, \sqrt{2}\, x_{i1} x_{i2}, x_{i2}^2, \sqrt{2}\, x_{i1}, \sqrt{2}\, x_{i2}]^T$

  - $\Psi(\mathbf{x_1}) = [1, 1, \sqrt{2}, 1, -\sqrt{2}, -\sqrt{2}]^T$
  - $\Psi(\mathbf{x_2}) = [1, 1, -\sqrt{2}, 1, -\sqrt{2}, \sqrt{2}]^T$
  - $\Psi(\mathbf{x_3}) = [1, 1, -\sqrt{2}, 1, \sqrt{2}, -\sqrt{2}]^T$
  - $\Psi(\mathbf{x_4}) = [1, 1, \sqrt{2}, 1, \sqrt{2}, \sqrt{2}]^T$

Example:

- $\Psi(\mathbf{x_1})^T\Psi(\mathbf{x_1}) =$

  $1 + 1 + 2 + 1 + 2 + 2 = 9$
- $\Psi(\mathbf{x_1})^T\Psi(\mathbf{x_3}) =$

  $1 + 1 - 2 + 1 - 2 + 2 = 1$

Example: XOR problem (continued)

- The matrix $\mathbb{K}$ is then

$$\mathbb{K} = \begin{bmatrix} 9 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 \\ 1 & 1 & 9 & 1 \\ 1 & 1 & 1 & 9 \end{bmatrix}$$

Example:

$\mathbb{K}(1,3) = [\mathbf{x}_1^T \mathbf{x}_3 + 1]^2$

$= [(-1,-1)^T(+1,-1) + 1)]^2$

$= (0+1)^2 = 1$

- The dual objective function $Q(\boldsymbol{\alpha})$ is

$$Q(\boldsymbol{\alpha}) = \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 - ½\,(9\alpha_1^2 - 2\alpha_1\alpha_2 - 2\alpha_1\alpha_3 + 2\alpha_1\alpha_4 + 9\alpha_2^2 + 2\alpha_2\alpha_3 - 2\alpha_2\alpha_4 + 9\alpha_3^2 - 2\alpha_3\alpha_4 + 9\alpha_4^2)$$

See Equation on page 37.

- Taking the derivative of $Q(\boldsymbol{\alpha})$ with respect to α_1, α_2, α_3 and α_4:

$$\begin{aligned} 9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 &= 1 \\ -\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 &= 1 \\ -\alpha_1 + \alpha_2 + 9\,\alpha_3 - \alpha_4 &= 1 \\ \alpha_1 - \alpha_2 - \alpha_3 + 9\,\alpha_4 &= 1 \end{aligned}$$

The XOR problem (continued)

- The solution: $\alpha_{*,1} = \alpha_{*,2} = \alpha_{*,3} = \alpha_{*,4} = 1/8$
- The optimum value of $Q(\boldsymbol{\alpha})$ is $Q(\boldsymbol{\alpha}_*) = 1/8 + 1/8 + 1/8 + 1/8 - 1/2(9/64 \ldots\ldots + 9/64) = 1/4$.
- This optimal value of the dual QP is equal to the optimal value of the primal QP, i.e.

$$Q(\boldsymbol{\alpha}_*) = 1/4 = 1/2\,\|\mathbf{w}_*\|^2 \quad \text{hence} \quad \|\mathbf{w}_*\| = 1/\sqrt{2}$$

- The optimum weight vector is

$$\mathbf{w}_* = 1/8\,[-\Psi(\mathbf{x}_1) + \Psi(\mathbf{x}_2) + \Psi(\mathbf{x}_3) - \Psi(\mathbf{x}_4)]$$

See Equation on page 35

$$= [0,\ 0,\ -1/\sqrt{2},\ 0,\ 0,\ 0]^T$$

- The bias b is the first element of $\mathbf{w}_*$ and it is equal to 0.
- The optimal hyperplane is $\mathbf{w}_*^T\,\Psi(\mathbf{x}) = 0$

4. Building SVM for pattern recognition

The XOR problem (continued)

- The optimal hyperplane is $\mathbf{w_0}^T \Psi(\mathbf{x}) = 0$

  $= [0,\ 0,\ -1/\sqrt{2}, 0,\ 0,\ 0]^T\ [1,\ x_1^2, \sqrt{2}\,x_1x_2, x_2^2, \sqrt{2}\,x_1, \sqrt{2}\,x_2]$

  which reduces to $-x_1 x_2 = 0$.

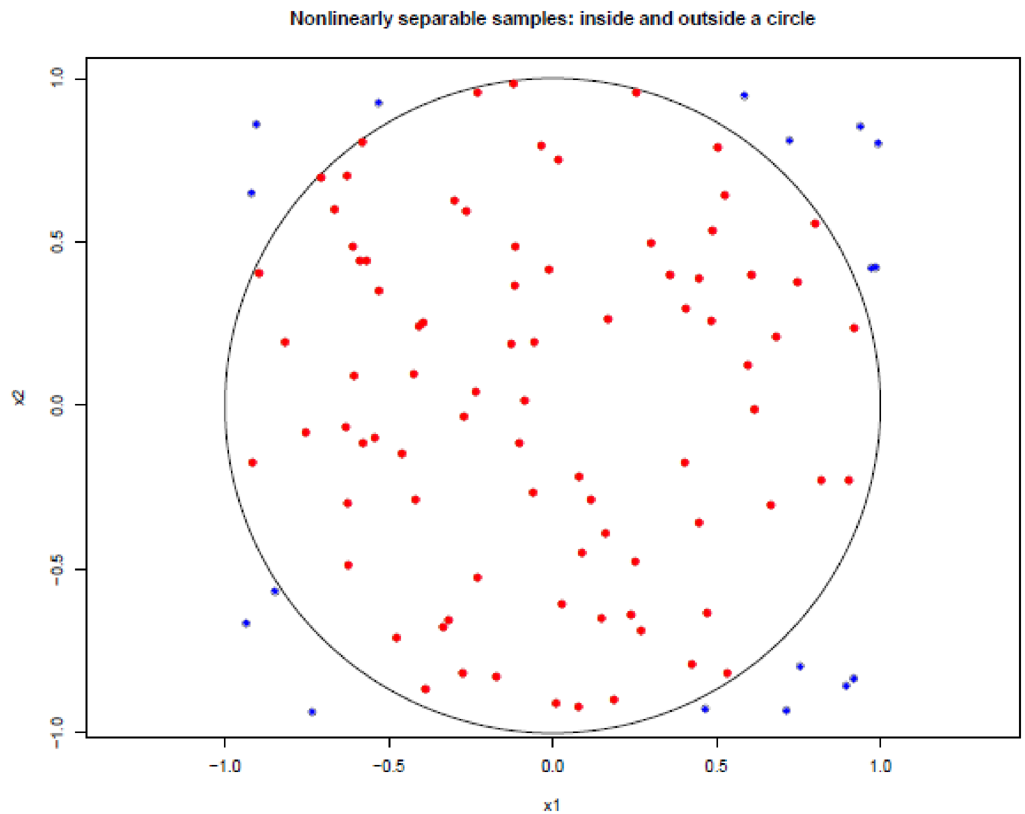
- Hence the outputs of the SVM can be summarised as follows:

XOR problem:		
Input vector **x**	Desired response d	Output $-x_1 x_2$
(-1,-1)	-1	-1
(-1,+1)	+1	+1
(+1,-1)	+1	+1
(+1,+1)	-1	-1

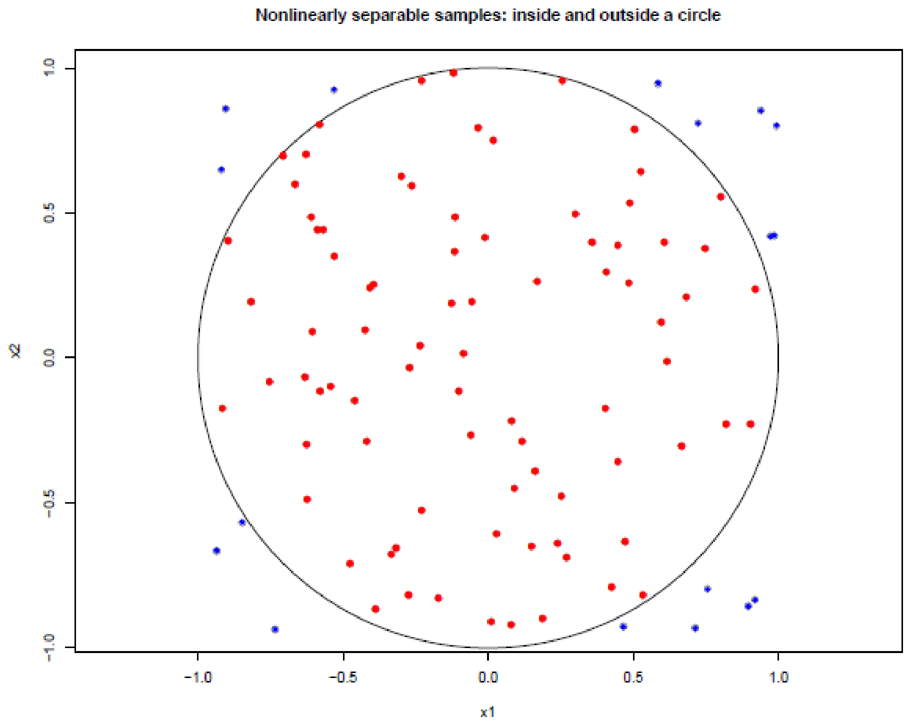
4. Building SVM for pattern recognition

Another example:

- Data in the original 2-dimensional space are not linearly separable.
- Suppose samples outside the circle are assigned Class A with $d_i = +1$, those on the circle or inside it Class B $d_i = -1$.
- Let the center of the circle be (a,b) and its radius be r
- Class A: $(x_1 - a)^2 + (x_2 - b)^2 > r^2$
- Class B: $(x_1 - a)^2 + (x_2 - b)^2 \leq r^2$



4. Building SVM for pattern recognition



Another example (continued):

- Class A: $(x_1 - a)^2 + (x_2 - b)^2 > r^2$
- Class B: $(x_1 - a)^2 + (x_2 - b)^2 \leq r^2$
- Apply polynomial kernel:

$$K(\mathbf{x}, \mathbf{x}_i) = 1 + x_1^2 x_{i1}^2 + 2x_1 x_2 x_{i1} x_{i2} + x_2^2 x_{i2}^2 + 2x_1 x_{i1} + 2x_2 x_{i2}$$

$$\Phi(\mathbf{x}) = \begin{pmatrix} 1 \\ x_1^2 \\ \sqrt{2} x_1 x_2 \\ x_2^2 \\ \sqrt{2} x_1 \\ \sqrt{2} x_2 \end{pmatrix}$$

Let:

$$\begin{aligned} w_0 &= a^2 + b^2 - r^2 \\ w_1 &= 1 \\ w_2 &= 0 \\ w_3 &= 1 \\ w_4 &= -\sqrt{2} a \\ w_5 &= -\sqrt{2} b \end{aligned}$$

Samples can be separated by the hyperplane

$$\mathbf{w}^T \Phi(\mathbf{x}) = 0$$

4. Building SVM for pattern recognition

Another example (continued):

- Class A: $(x_1 - a)^2 + (x_2 - b)^2 > r^2$
- Class B: $(x_1 - a)^2 + (x_2 - b)^2 \leq r^2$

$$\Phi(\mathbf{x}) = \begin{pmatrix} 1 \\ x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \end{pmatrix}$$

Let:

$$\begin{aligned} w_0 &= a^2 + b^2 - r^2 \\ w_1 &= 1 \\ w_2 &= 0 \\ w_3 &= 1 \\ w_4 &= -\sqrt{2}a \\ w_5 &= -\sqrt{2}b \end{aligned}$$

Samples can be separated by the hyperplane

$$\mathbf{w}^T \Phi(\mathbf{x}) = 0$$

$$a^2 + b^2 - r^2 + x_1^2 + 0 + x_2^2 - 2ax_1 - 2bx_2 = 0$$

$$\updownarrow$$

$$x_1^2 - 2ax_1 + a^2 + x_2^2 - 2bx_2 + b^2 - r^2 = 0$$

$$\updownarrow$$

$$(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$$

<u>Decision:</u>

If $\mathbf{w}^T \Phi(\mathbf{x}) > 0$, then class A, else class B

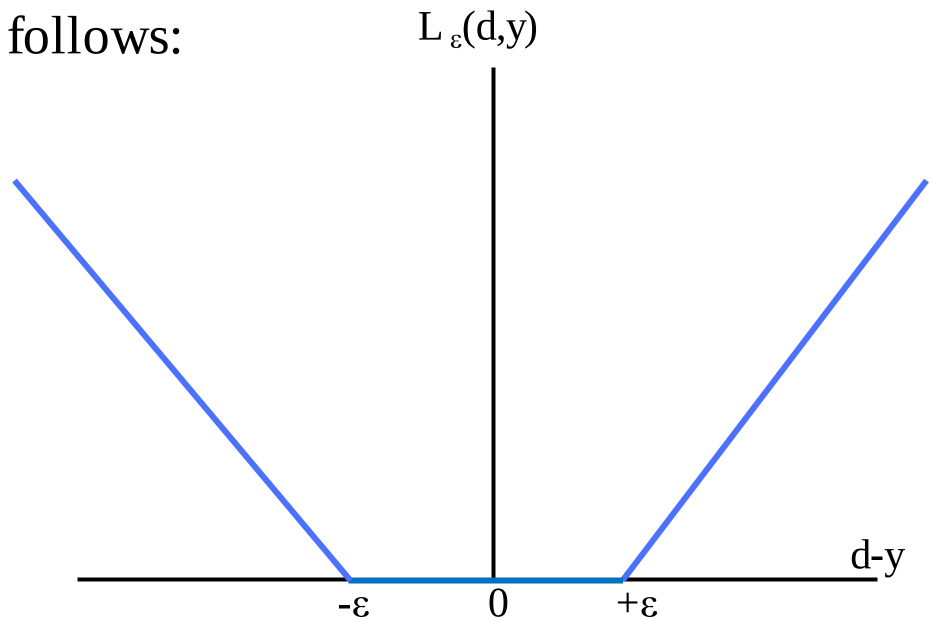
5. SVM for nonlinear regression

Definition: <u>ε-insensitive loss function</u>

- Let d be the desired response and y its estimated value.
- The <u>ε-insensitive loss function</u> is defined as follows:

$$L_\varepsilon(d, y) = |d - y| - \varepsilon \ \text{ if } \ |d - y| \geq \varepsilon,$$

$$0 \text{ otherwise}$$



- Expressing $L_\varepsilon(d, y)$ as a linear program:

minimise $\xi + \xi'$

subject to $d - y \leq \varepsilon + \xi$

$y - d \leq \varepsilon + \xi'$

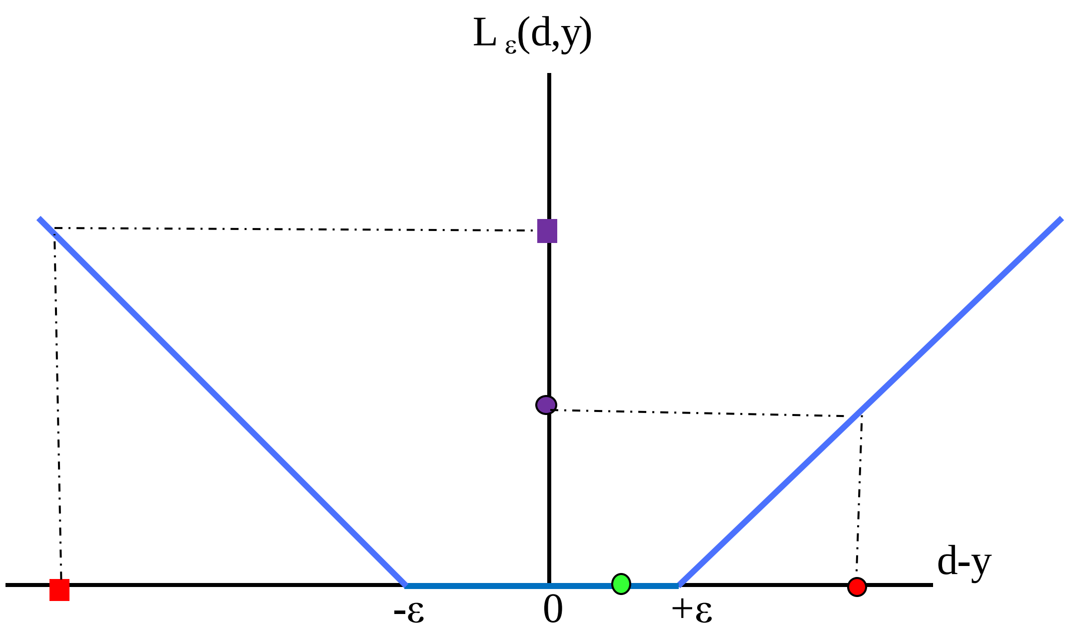
$\xi', \ \xi \geq 0$

5. SVM for nonlinear regression

ε-insensitive loss function (continued)

- <u>Note:</u> if $d - y \geq \varepsilon$, then let $\xi = d - y - \varepsilon \geq 0$, $\xi' = 0$ and $y - d \leq -\varepsilon \leq \varepsilon$.

  if $d - y \leq -\varepsilon$, then let $\xi' = y - d - \varepsilon \geq 0$ and $\xi = 0$.



Example 1:
$d-y = 2\varepsilon$
$\xi = d - y - \varepsilon = 2\varepsilon - \varepsilon = \varepsilon$ (this is the penalty)
$\xi' = 0$

Example 2:
$d - y = -3\varepsilon$
$\xi' = y - d - \varepsilon = 2\varepsilon$ (this is the penalty)
$\xi = 0$

Example 3:
$d-y = \frac{1}{2}\varepsilon$
$\xi = \xi' = 0$ (no penalty)

In all three examples, check that:

$d - y \leq \varepsilon + \xi$
$y - d \leq \varepsilon + \xi'$

5. SVM for nonlinear regression

- For nonlinear regression problem, given the training data $\{(\mathbf{x_i}, d_i)\}$ i = 1,2,…., N , where $\mathbf{x_i}$ is a sample value of the input vector $\mathbf{x}$ and d_i is the corresponding value of the model output d, we minimise the following QP:

$$\text{minimise } \Phi(\mathbf{w}, \boldsymbol{\xi}, \boldsymbol{\xi}') = C\left(\sum_{i=1}^{N} (\xi_i + \xi_i')\right) + \tfrac{1}{2}\, \mathbf{w}^T \mathbf{w}$$

$$\text{subject to } \quad d_i - \mathbf{w}^T\, \Psi(\mathbf{x_i}) \le \varepsilon + \xi_i$$

$$\mathbf{w}^T\, \Psi(\mathbf{x_i}) - d_i \le \varepsilon + \xi_i'$$

$$\xi_i',\ \xi_i \ge 0, \quad i = 1,2, \ldots, N$$

- C is a positive parameter selected by the user.

5. SVM for nonlinear regression

- The dual of the quadratic problem is

$$\text{maximise } Q(\boldsymbol{\alpha}_{\mathbf{i}}, \boldsymbol{\alpha}_{\mathbf{i}}') = \sum_{i=1}^{N} d_i(\alpha_i - \alpha_i') - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i')$$

$$- \tfrac{1}{2} \sum_{i=1}^{N} \sum_{i=1}^{N} (\alpha_i - \alpha_i')(\alpha_j - \alpha_j') K(\mathbf{x_i}, \mathbf{x_j})$$

$$\text{subject to } \sum_{i=1}^{N} (\alpha_i - \alpha_i') = 0$$

$$0 \le \alpha_i \le C, \quad i = 1,2, \ldots., N$$

$$0 \le \alpha_i' \le C, \quad i = 1,2, \ldots., N$$

- The parameter ε and C must be tuned simultaneously.

- Regression is intrinsically more difficult than pattern classification.

6. SVM for handling imbalanced data

- The paper "A synthetic informative minority over-sampling (SIMO) algorithm leveraging support vector machine to enhance learning from imbalanced datasets", S. Piri, D. Delen, T. Liu, Decision Support Systems 106 (2018) 15-29 presents methods to oversample minority examples in the data by using SVM.

- Idea: pay more attention to minority examples that are close to the SVM decision boundary.

6. Calculate the Euclidean distance of minority data points form D_B

$$Euc\_D(x^{k+}) = \frac{|\sum_{t=1}^{m} w_t x_t^{k+} + b|}{\sqrt{\sum_{t=1}^{m} {w_t}^2}}$$

- SIMO algorithm: informative minority examples close to the SVM decision boundary are over-sampled.

- W-SIMO: weighted SIMO, incorrectly classified informative minority examples are over-sampled with a higher-degree compared to the correctly classified informative minority examples.

6. SVM for handling imbalanced data

•

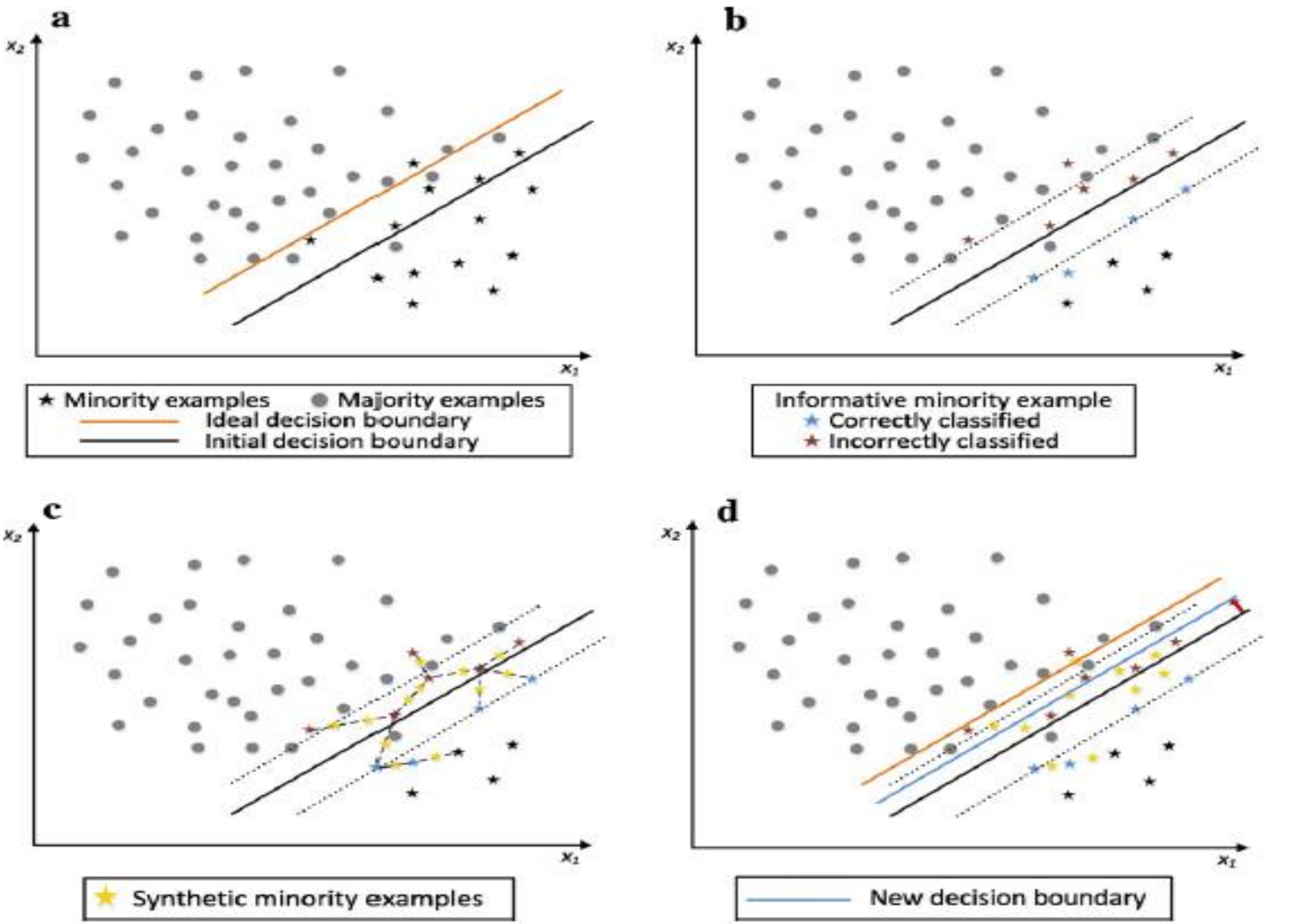


Fig. 2. W-SIMO algorithm mechanism (simplified).

6. SVM for handling imbalanced data

- Experiments on 15 data sets:

Table 2
Benchmark datasets characteristics.

Dataset	Minority class	Majority class	# of variables	# of records	Imbalance ratio
Liver disorders (liver)	"1"	"2"	7	345	1:1.38
Ionosphere	Bad	Good	34	351	1:1.79
Pima Indians Diabetes (Pima)	"1"	"0"	8	768	1:1.87
Breast Cancer Wisconsin Original (BreastCO)	Malignant	Benign	10	699	1:1.91
Iris	Versicolor	All other	5	150	1:2
Yeast	NUC	All other	8	1484	1:2.6
Statlog Vehicle Silhouettes (vehicle)	Van	All other	18	846	1:3.25
Contraceptive Method Choice (CMC)	Long-term	All other	9	1473	1:3.42
Breast Cancer Wisconsin_20% (BreastC20)	Malignant	Benign	10	699	1:3.91
Connectionist Bench_Vowel Recognition (vowel)	"0" & "1"	All other	11	990	1:4.5
Ecoli	pp	All other	8	336	1:5.46
Libras Movement_12 (Libras12)	"1" & "2"	All other	91	360	1:5.88
Libras Movement_34 (Libras34)	"3" & "4"	All other	91	360	1:6.34
Glass identification (glass)	"7"	All other	9	214	1:6.38
Breast Cancer Wisconsin_10% (BreastC10)	Malignant	Benign	10	699	1:8.9

6. SVM for handling imbalanced data

- Experiments on 15 data sets:

Table 6
Overall ranking on linear SVM.

Approach	G mean	AUC
W-SIMO	1.1	1.1
SIMO	1.9	1.9
SMOTE-IPF	4.6	4.5
Cluster SMOTE	5.3	5.0
Cost sensitive	5.7	6.1
SMOTE	6.3	6.4
Safe level SMOTE	6.3	6.7
Under sampling	7.3	7.1
BorSMOTE	8.1	8.1
Original data	8.6	8.1

- G mean = sqrt(TPR × TNR)

Reference

- S. Haykin, Neural Networks A comprehensive foundation, second edition, 1999, Prentice Hall.
- Slide 3 data set is from J.D. Kelleher et al, Fundamentals of Machine Learning for Predictive Data Analytics
- Also visit http://www.support-vector-machines.org Q325.5 Cri 2000

  and http://www.csie.ntu.edu.tw/~cjlin/libsvm/index.html (a library for support vector machines).