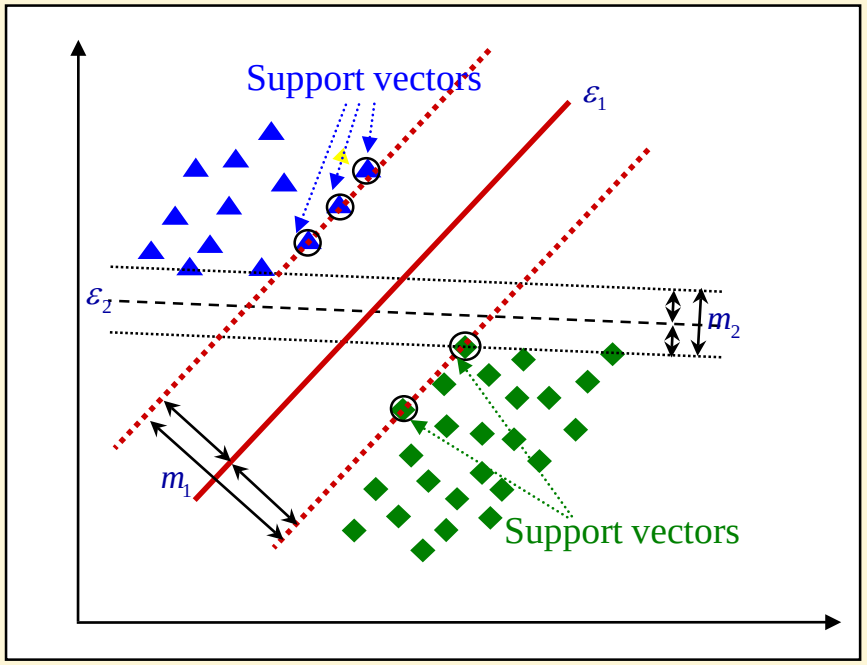
Support Vector Machines: Linear Classification

Margin and support vectors

- Consider a problem of two linearly separable classes.
- The perceptron algorithm terminates on finding any discriminative hyperplane between the two classes of the patterns.
- All possible solutions are considered as equivalent. There is no provision for characterising some of these solutions as preferable.
- Is it possible to formulate a criterion for choosing among the possible solutions?



Two discriminating straight lines in a two dimensional space. Note the margin (m_1 και m_2 respectively in the two cases). The maximum margin is achieved in the case of the straight line ε_1.

MAXIMUM MARGIN CRITERION:

We seek to find a hyperplane which separates the classes completely, so that:

- This hyperplane is equidistant from the closest patterns of the two classes. The distance between the hyperplance and the closest pattern is called the margin
- The margin is the maximum possible

The patterns that are closest to the hyperplane (and therefore their distance from the hyperplane is equal to the margin) are called support vectors.

- Equation of discriminating hyperplane: $\mathbf{w} \cdot \mathbf{x} + w_0 = 0$
- Distance of vector $\hat{\mathbf{x}}$ from the hyperplane: $\dfrac{\left|(\mathbf{w} \cdot \hat{\mathbf{x}}) + w_0\right|}{\|\mathbf{w}\|}$
- Obviously, if we multiply all synaptic weights by a coefficient ρ, the discriminating hyperplane does not change. We can therefore select ρ, so that the following condition holds for the support vectors $\hat{\mathbf{x}}_s$:

$$\mathbf{w} \cdot \hat{\mathbf{x}}_s + w_0 = \begin{cases} +1, & \hat{\mathbf{x}}_s \in C_1 \\ -1, & \hat{\mathbf{x}}_s \in C_2 \end{cases}$$

- With this assumption, the margin is equal to: $\dfrac{1}{\|\mathbf{w}\|} + \dfrac{1}{\|\mathbf{w}\|} = \dfrac{2}{\|\mathbf{w}\|}$
- Moreover:

$$\mathbf{w} \cdot \mathbf{x}_i + w_0 \geq 1 \quad \forall \mathbf{x}_i \in C_1$$
$$\mathbf{w} \cdot \mathbf{x}_i + w_0 \leq -1 \quad \forall \mathbf{x}_i \in C_2$$

- Equivalently: $t_i(\mathbf{w} \cdot \mathbf{x}_i + w_0) \geq 1$ (outputs +1 and -1 for the 2 classes)

$$t_i = 1 \quad \forall \mathbf{x}_i \in C_1$$
$$t_i = -1 \quad \forall \mathbf{x}_i \in C_2$$

We arrive at the following optimization problem:

- Minimization of: $\frac{\|\mathbf{w}\|^2}{2}$

- Under the constraints:

$$t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1$$
$$t_i = 1 \quad \forall \mathbf{x}_i \in C_1$$
$$t_i = -1 \quad \forall \mathbf{x}_i \in C_2$$

(this is a quadratic optimization problem under linear inequality constraints)

We introduce Lagrange multipliers.
The Lagrangian:

$$L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{P} \lambda_i (t_i \cdot ((\mathbf{x}_i \cdot \mathbf{w}) + w_0) - 1)$$

- **Unique solution:**
  There is a unique hyperplane which solves the problem, because
  - The cost function is convex
  - The linear inequality constraints always form a convex domain containing the solution.

$$L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{P} \lambda_i (t_i \cdot ((\mathbf{x}_i \cdot \mathbf{w}) + w_0) - 1)$$

The solution is furnished by the conditions:

KARUSH-KUHN-TUCKER CONDITIONS

$$\frac{\partial}{\partial w_0} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = 0$$

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = 0$$

$$\lambda_i \geq 0, \quad i = 1, 2, ..., P$$

$$\lambda_i \left[t_i (\mathbf{w} \cdot \mathbf{x}_i + w_0) - 1 \right] = 0, \quad i = 1, 2, ..., P$$

i.e.:

$$\sum_{i=1}^{P} \lambda_i t_i = 0$$

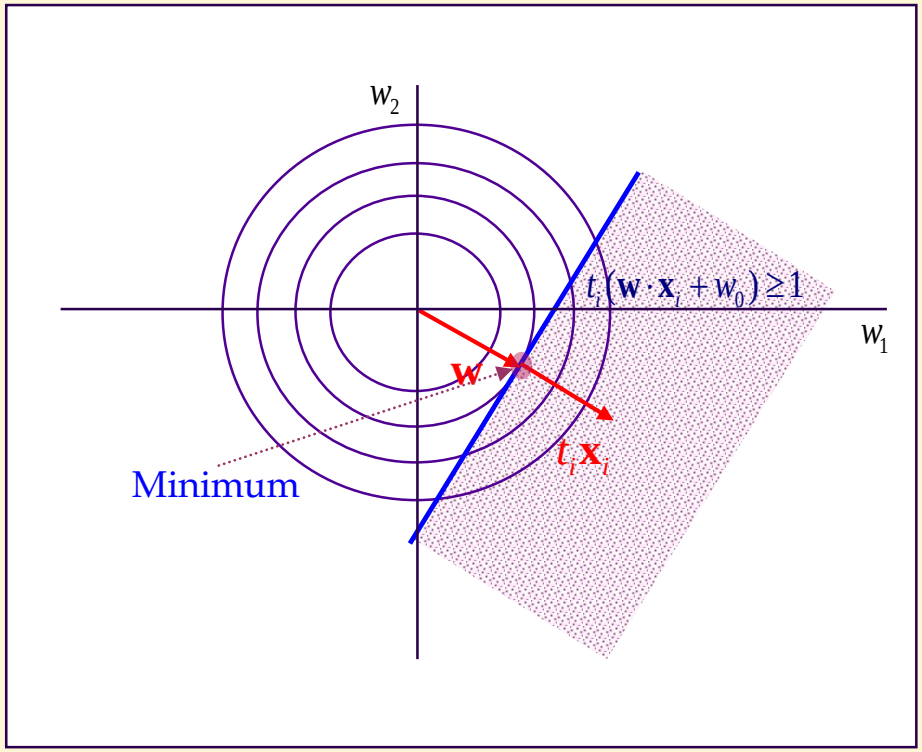
$$\mathbf{w} = \sum_{i=1}^{P} \lambda_i t_i \mathbf{x}_i$$

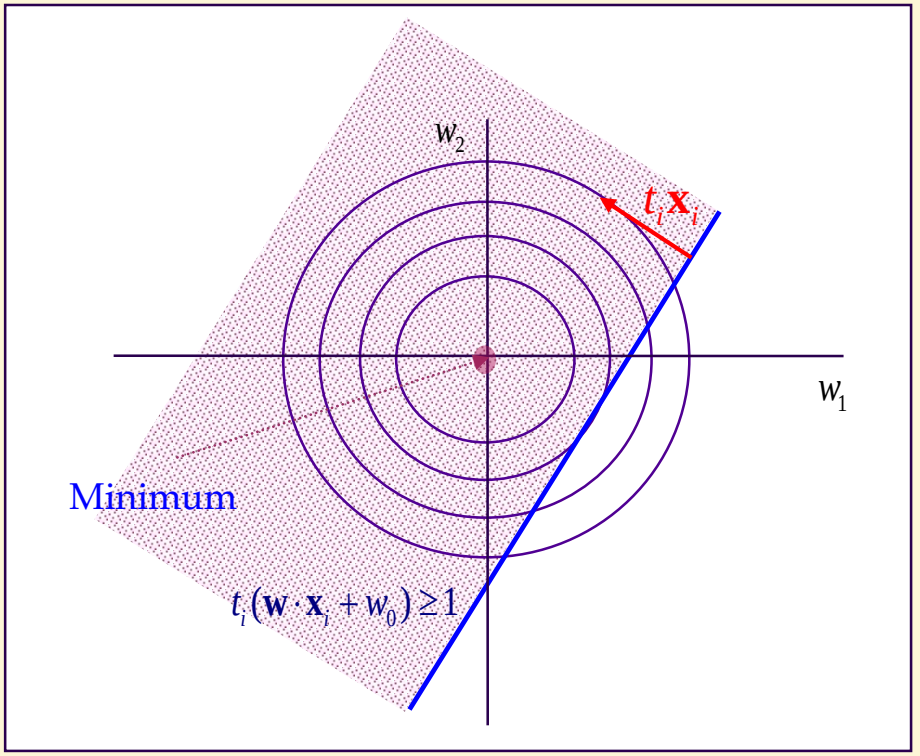
$$\lambda_i \geq 0, \quad i = 1, 2, ..., P$$

$$\lambda_i \left[t_i (\mathbf{w} \cdot \mathbf{x}_i + w_0) - 1 \right] = 0, \quad i = 1, 2, ..., P$$

Notice the similarity with the perceptron rule! Remember the dual variables!

Schematically:





1st case: Active constraint $(\lambda_i > 0)$
Alignment of vectors $\mathbf{w}$ *and* $t_i\mathbf{x}_i$ *with a positive coefficient of proportionality.*

2nd case: Inactive constraint. The constraint does not affect the minimum and therefore $\lambda_i = 0$

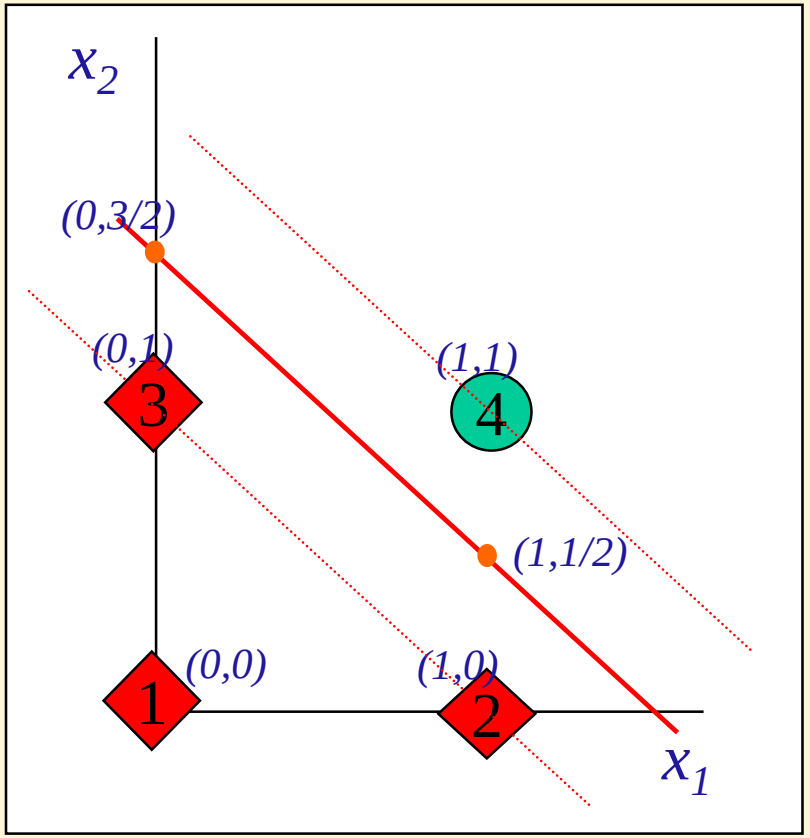
Obviously, the constraints which are active at the solution point correspond to the support vectors, because at that point the following holds:

$$t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) = 1 \Rightarrow \mathbf{w}\cdot\mathbf{x}_i + w_0 = \pm 1$$

Example: AND problem (1)

A/A	x_1	x_2	t
1	0	0	-1
2	1	0	-1
3	0	1	-1
4	1	1	1

Inputs and desired outputs for the AND problem



- Pattern number 4 is the sole member of its class. Necessarily, it is a support vector.
- It is easy to see by the maximum margin requirement that patterns number 2 and 3 are also support vectors.
- It follows that the optimal discriminating straight line passes through the points (0,3/2) και (1,1/2) on the (x_1, x_2) plane.

Example: AND problem (2)

- Let us confirm the Karush-Kuhn-Tucker conditions
- The discriminating straight line passes through the points (0,3/2) και (1,1/2):

$$\tfrac{3}{2} w_2 + w_0 = 0 \Rightarrow w_2 = -\tfrac{2}{3} w_0$$

$$w_1 + \tfrac{1}{2} w_2 + w_0 = 0 \Rightarrow w_1 = -\tfrac{2}{3} w_0$$

- Equation of the discriminating line:

$$(-\tfrac{2}{3} x_1 - \tfrac{2}{3} x_2 + 1) w_0 = 0$$

- The pattern (1,1) is a support vector, therefore:

$$(-\tfrac{2}{3} \cdot 1 - \tfrac{2}{3} \cdot 1 + 1) w_0 = 1 \Rightarrow w_0 = -3$$

- It follows that:

$$w_1 = 2, \quad w_2 = 2, \quad w_0 = -3$$

- We can immediately confirm that the patterns (1,0) and (0,1) are support vectors:

$$w_1 \cdot 1 + w_2 \cdot 0 + w_0 = 2 \cdot 1 - 3 = -1$$

$$w_1 \cdot 0 + w_2 \cdot 1 + w_0 = 2 \cdot 1 - 3 = -1$$

Example: AND problem (3)

- Let us examine the Lagrange multipliers:
- The pattern (0,0) is not a support vector, therefore $\lambda_1 = 0$
- We expect that the remaining Lagrange multipliers, corresponding to support vectors, are positive. We can find them using the conditions:

$$\sum_{i=1}^{4} \lambda_i t_i = 0, \qquad \mathbf{w} = \sum_{i=1}^{4} \lambda_i t_i \mathbf{x}_i$$

- Therefore:

$$\lambda_4 - \lambda_2 - \lambda_3 = 0$$

$$\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \lambda_4 \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \lambda_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \lambda_3 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Solving this system of linear equations we get:

$$\lambda_4 = 4, \lambda_2 = 2, \lambda_3 = 2$$

- The Lagrange multipliers are positive, as expected.

The dual problem

- In the majority of cases it is not possible to find analytically the solution indicated by the Karush-Kuhn-Tucker conditions. In these cases we are obliged to solve the optimization problem iteratively using appropriate algorithms
- However, we can exploit the Karush-Kuhn-Tucker conditions in order to reformulate the problem in terms of the dual variables only. Let us introduce the conditions $\mathbf{w} = \sum_{i=1}^{P} \lambda_i t_i \mathbf{x}_i, \quad \sum_{i=1}^{P} \lambda_i t_i = 0$ into the original Lagrangian:

$$L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \tfrac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{P} \lambda_i (t_i \cdot ((\mathbf{x}_i \cdot \mathbf{w}) + w_0) - 1) =$$

$$\tfrac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{P} \lambda_i t_i \cdot (\mathbf{x}_i \cdot \mathbf{w}) - w_0 \sum_{i=1}^{P} \lambda_i t_i + \sum_{i=1}^{P} \lambda_i$$

0

This helps us eliminate the synaptic weights as follows:

$$\tfrac{1}{2}\|\mathbf{w}\|^2 = \tfrac{1}{2}\mathbf{w}\cdot\mathbf{w} = \tfrac{1}{2}\left(\sum_{i=1}^{P}\lambda_i t_i \mathbf{x}_i\right)\cdot\left(\sum_{j=1}^{P}\lambda_j t_j \mathbf{x}_j\right) = \tfrac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j t_i t_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)$$

$$\sum_{i=1}^{P}\lambda_i t_i\cdot(\mathbf{x}_i\cdot\mathbf{w}) = \sum_{i=1}^{P}\lambda_i t_i\cdot(\sum_{j=1}^{P}\lambda_j t_j \mathbf{x}_i\cdot\mathbf{x}_j) = \sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j t_i t_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)$$

whereupon the Lagrangian becomes: $\sum_{i=1}^{P}\lambda_i - \tfrac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j t_i t_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)$

Dual Problem Formulation

Maximize, with respect to λ_i : $L_D = \sum_{i=1}^{P}\lambda_i - \tfrac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j t_i t_j\left(\mathbf{x}_i\cdot\mathbf{x}_j\right)$

under the constraints: $\sum_{i=1}^{P}\lambda_i t_i = 0 \qquad \lambda_i \geq 0, \quad i = 1, 2, ..., P$

REMARK 1: The inequality constraints are now simpler, since they are mutually orthogonal. However, the cost function is more complicated: its contours changed from hyperspheres to hyperellipses.

REMARK 2: The dual formulation helps us devise iterative algorithms for finding the support vectors and the synaptic weights.

REMARK 3: In the dual problem, the threshold w_0 is absent from the Lagrangian. Once the λ_i s are determined, the threshold is found using any support vector $\mathbf{x}_s$ via the relation:

$$t_s(\mathbf{w}\cdot\mathbf{x}_s + w_0) = 1 \Rightarrow w_0 = t_s - \sum_i \lambda_i t_i (\mathbf{x}_s \cdot \mathbf{x}_i)$$

REMARK 4: After completion of the training phase, any new pattern $\mathbf{x}_{new}$ is classified as follows:

$$\sum_i \lambda_i t_i (\mathbf{x}_{new}\cdot\mathbf{x}_i) + w_0 > 0 \Rightarrow \mathbf{x}_{new} \in C_1$$

$$\sum_i \lambda_i t_i (\mathbf{x}_{new}\cdot\mathbf{x}_i) + w_0 < 0 \Rightarrow \mathbf{x}_{new} \in C_2$$

The patterns are appear in inner product form both in the Lagrangian and in the classification rule. This is very important because it will help us ***generalize the SVM formulation for the classification of non-linearly separable data.***

Example: AND problem (dual formulation)

Dual problem

Lagrangian:

$$L_D = \sum_{i=1}^{P} \lambda_i - \tfrac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j t_i t_j \left(\mathbf{x}_i \cdot \mathbf{x}_j \right)$$

A/A	x_1	x_2	t
1	0	0	-1
2	1	0	-1
3	0	1	-1
4	1	1	1

$$t_2^{\,2}(\mathbf{x}_2 \cdot \mathbf{x}_2) = (-1)(-1)(+1) = 1$$

$$t_3^{\,2}(\mathbf{x}_3 \cdot \mathbf{x}_3) = (-1)(-1)(+1) = 1$$

$$t_4^{\,2}(\mathbf{x}_4 \cdot \mathbf{x}_4) = (+1)(+1)(+2) = 2$$

$$t_2 t_4(\mathbf{x}_2 \cdot \mathbf{x}_4) = (-1)(+1)(1) = -1$$

$$t_3 t_4(\mathbf{x}_3 \cdot \mathbf{x}_4) = (-1)(+1)(1) = -1$$

$$t_1^{\,2}(\mathbf{x}_1 \cdot \mathbf{x}_1) = 0$$

$$t_1 t_2(\mathbf{x}_1 \cdot \mathbf{x}_2) = t_1 t_3(\mathbf{x}_1 \cdot \mathbf{x}_3) = t_1 t_4(\mathbf{x}_1 \cdot \mathbf{x}_4) = t_2 t_3(\mathbf{x}_2 \cdot \mathbf{x}_3) = 0$$

Example: AND problem (dual formulation) (2)

Maximize: $L_D = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 - \frac{1}{2}(\lambda_2^2 + \lambda_3^2 + 2\lambda_4^2 - 2\lambda_2\lambda_4 - 2\lambda_3\lambda_4)$

Constraint: $-\lambda_1 - \lambda_2 - \lambda_3 + \lambda_4 = 0$

Since all relations are symmetric with respect to λ_2 and λ_3, we seek a solution with $\lambda_2 = \lambda_3$

$\Rightarrow$ Maximize: $L_D = \lambda_1 + 2\lambda_2 + \lambda_4 - \lambda_2^2 - \lambda_4^2 + 2\lambda_2\lambda_4$

Constraint: $\lambda_4 = \lambda_1 + 2\lambda_2, \ \lambda_2 = \lambda_3$

Let us examine the following cases:

A) All patterns are support vectors

B) Pattern number 1 is not a support vector $\lambda_1 = 0$

Γ) Patterns number 2, 3 are not support vectors $\lambda_2 = \lambda_3 = 0$

Δ) Pattern number 4 is not a support vector $\lambda_4 = 0$

All other cases are trivial and lead easily to

$\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0$

Example: AND problem (dual formulation) (3)

CASE A) All patterns are support vectors

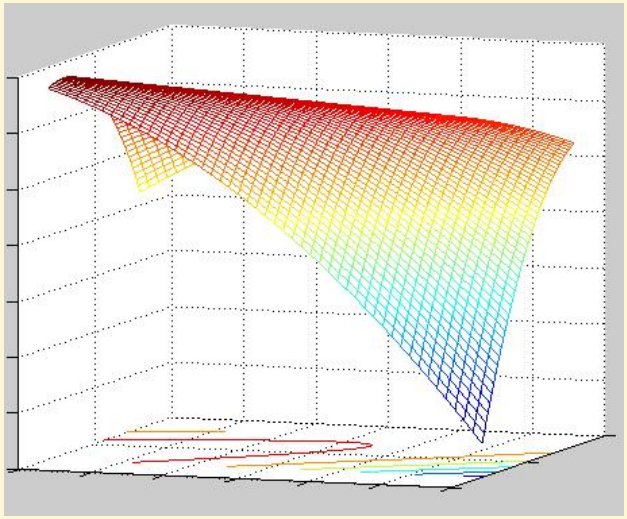
Using the constraint, we eliminate λ_1

$$\Rightarrow \qquad \text{Maximize: } L_D = 2\lambda_4 - (\lambda_2 - \lambda_4)^2$$

$$\frac{\partial L_D}{\partial \lambda_2} = -2(\lambda_2 - \lambda_4) = 0 \Rightarrow \lambda_2 - \lambda_4 = 0$$

$$\frac{\partial L_D}{\partial \lambda_4} = 2 + 2(\lambda_2 - \lambda_4) = 0 \Rightarrow \lambda_2 - \lambda_4 = -1$$

We have reached an absurd conclusion: There is no maximum, therefore there is no solution with all the patterns as support vectors.



Graphical representation of L_D*.*
There is no maximum.

Example: AND problem (dual formulation) (4)

CASE B) Pattern #1 is not a support vector: $\lambda_1 = 0$

$$\Rightarrow \quad \text{Maximize: } L_D = 2\lambda_2 + \lambda_4 - \lambda_2^{\,2} - \lambda_4^{\,2} + 2\lambda_2\lambda_4$$

$$\text{Constraints: } \lambda_4 = 2\lambda_2, \;\; \lambda_2 = \lambda_3$$

$$L_D = 4\lambda_2 - \lambda_2^{\,2}, \quad \frac{\partial L_D}{\partial \lambda_2} = 0 \Rightarrow \lambda_2 = 2, \text{ therefore } \lambda_3 = 2, \quad \lambda_4 = 2\lambda_2 = 4 \quad \text{and} \quad L_D = 4$$

CASE C) Patterns #2 and #3 are not support vectors: $\lambda_2 = \lambda_3 = 0$

$$\Rightarrow \quad \text{Maximize: } L_D = \lambda_1 + \lambda_4 - \lambda_4^{\,2}$$

$$\text{Constraint: } \lambda_4 = \lambda_1$$

$$L_D = 2\lambda_4 - \lambda_4^{\,2}, \quad \frac{\partial L_D}{\partial \lambda_4} = 0 \Rightarrow \lambda_4 = 1, \text{ therefore } \lambda_1 = 1, \quad L_D = 1$$

Example: AND problem (dual formulation) (4)

CASE D) Pattern 4 is not a support vector: $\lambda_4 = 0$

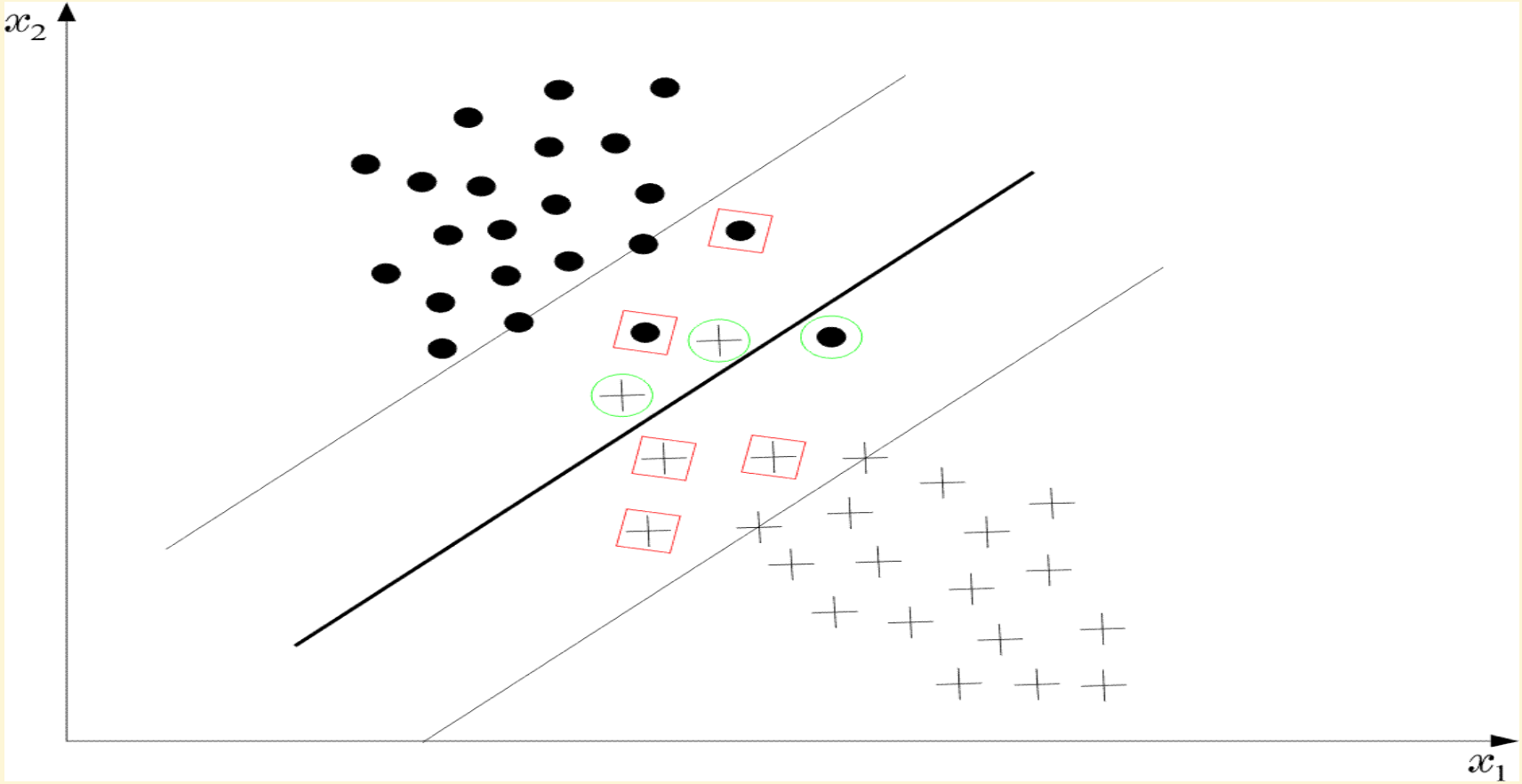
$$\Rightarrow \quad \text{Maximize: } L_D = \lambda_1 + 2\lambda_2 - \lambda_2^{\,2}$$

$$\text{Constraints: } \lambda_1 + 2\lambda_2 = 0, \;\; \lambda_2 = \lambda_3$$

$$L_D = -\lambda_2^{\,2}, \quad \frac{\partial L_D}{\partial \lambda_2} = 0 \Rightarrow \lambda_2 = 0, \text{ therefore } \lambda_1 = \lambda_3 = \lambda_4 = 0 \quad \text{and} \quad L_D = 0$$

Concluding, case (B) yields the maximum value of the dual Lagrangial. The solution entails patterns #2, #3 and #4 as support vectors. As expected, we obtain again what we had verified earlier as a solution of the Karush-Kuhn-Tucker conditions.

Linearly non-separable classes (1)



Linearly non-separable classes (2)

- The training patterns belong to one of the following classes:

1) Rightly classified patterns outside the margin:

$$t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1$$

2) Rightly classified patterns inside the margin:

$$0 \leq t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) < 1$$

3) Wrongly classified patterns:

$$t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) < 0$$

Linearly non-separable classes (3)

- We can unify all cases as follows:

$$t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1 - \xi_i$$

1) $\rightarrow \xi_i = 0$
2) $\rightarrow 0 < \xi_i \leq 1$
3) $\rightarrow 1 < \xi_i$

- ξ_i : Auxiliary (slack) variables.
- They are non-negative.
- For correctly classified patterns outside the margin, they are zero.
- For all other patters, they are positive.

Linearly non-separable classes (4)

➢ Optimization scheme:

- Maximize the margin
- Minimize the # number of patterns with $\xi_i > 0$

  Cost function:

$$E(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{P} I(\xi_i)$$

  where C is a constant and

$$I(\xi_i) = \begin{cases} 1 & \xi_i > 0 \\ 0 & \xi_i = 0 \end{cases}$$

- *I(.)* non-differentiable. In practice, we use the alternative:

$$E(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{P} \xi_i$$

Linearly non-separable classes (5)

Optimization problem

Minimize:

$$E(\mathbf{w}, \boldsymbol{\xi}) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{P}\xi_i$$

Under the constraints:

$$t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) \geq 1 - \xi_i, \quad i = 1, 2, \ldots, P$$

$$\xi_i \geq 0, \quad i = 1, 2, \ldots, P$$

Linearly non-separable classes (6)

- Lagrangian:

$$L(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{P}\xi_i - \sum_{i=1}^{P}\mu_i\xi_i - \sum_{i=1}^{P}\lambda_i[t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) - 1 + \xi_i]$$

- Karush-Kuhn-Tucker conditions:

(1) $\mathbf{w} = \sum_{i=1}^{P}\lambda_i t_i \mathbf{x}_i$

(2) $\sum_{i=1}^{P}\lambda_i t_i = 0$

(3) $C - \mu_i - \lambda_i = 0,\ i = 1, 2, ..., P$

(4) $\lambda_i[t_i(\mathbf{w}\cdot\mathbf{x}_i + w_0) - 1 + \xi_i] = 0, \quad i = 1, 2, ..., P$

(5) $\mu_i\xi_i = 0, \quad i = 1, 2, ..., P$

(6) $\mu_i, \lambda_i \geq 0, \quad i = 1, 2, ..., P$

Linearly non-separable classes (7)

$$L = \frac{1}{2}\|\underline{w}\|^2 + \sum_{i=1}^{P} \lambda_i t_i \mathbf{w}\cdot\mathbf{x}_i + \sum_{i=1}^{P} (C - \lambda_i - \mu_i)\xi_i - w_0 \sum_{i=1}^{P} \lambda_i t_i + \sum_{i=1}^{P} \lambda_i$$

Substitute **w** from eq. (1)

= 0 by eq. (3)

= 0 by eq. (2)

- **The dual problem:**

  Maximize, with respect to **λ:**

$$\sum_{i=1}^{P} \lambda_i - \frac{1}{2}\sum_{i,j}^{P} \lambda_i \lambda_j t_i t_j \mathbf{x_i}\mathbf{x_j}$$

  under the constraints:

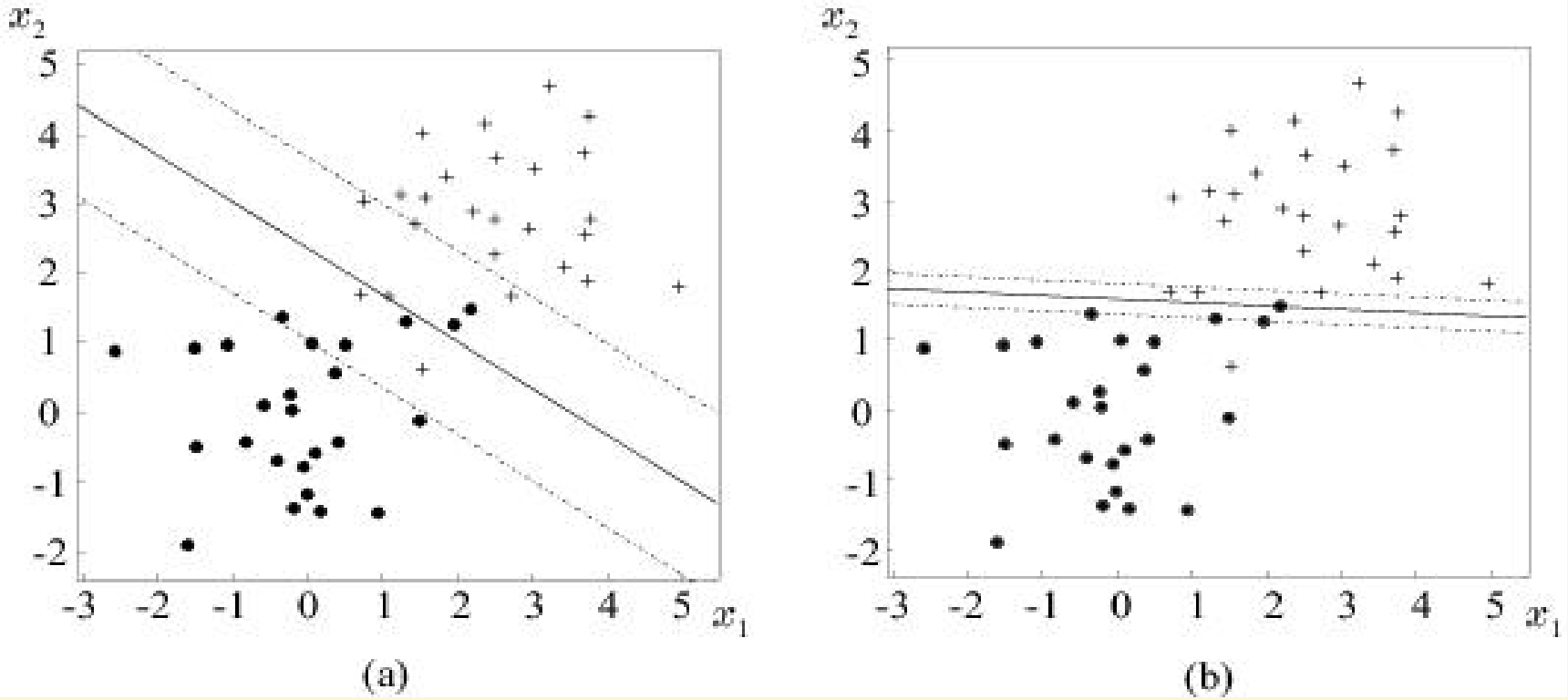
$$0 \le \lambda_i \le C, \quad i = 1, 2, ..., P$$

$$\sum_{i=1}^{P} \lambda_i t_i = 0$$

- Remark: The only difference with the case of linearly separable classes is the presence of C in the constraints.

Linearly non-separable classes (8)

- Example:



- Observe the influence of C (0,2 left, 1000 right).

Training algorithms

Problem: High computational cost. We employ techniques that break down the problem into smaller ones:

We initialize the process using a subset of the training set (working set). We optimize the dual Lagrangian for this subset and find the support vectors.

- The support vectors remain in the working set. The remaining vectors are substituted by other patterns not belonging to the working set. These are the patterns that violate the KKT conditions most severely.
- The process is repeated many times and guarantees the ongoing descent of the cost function.
- Platt's algorithm uses working sets with just 2 patterns. Optimization within the working set can be done analytically.

Multi-class problems

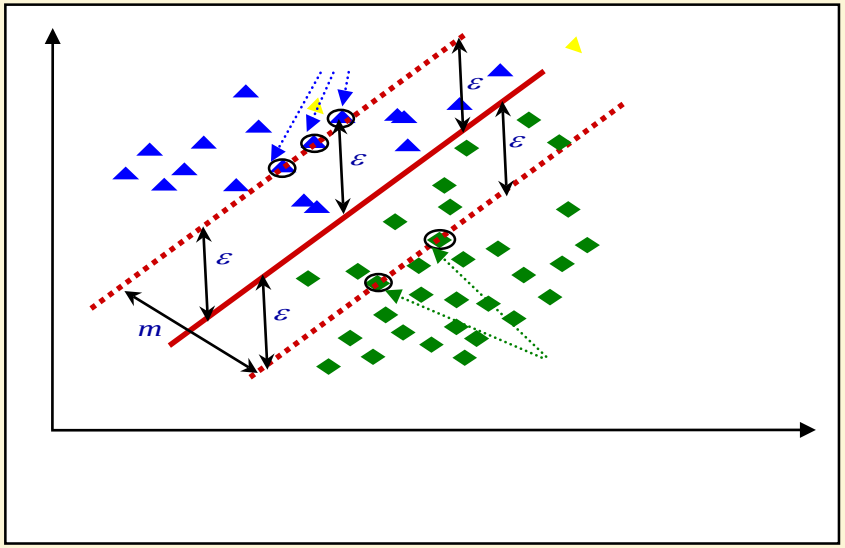
- The most popular practice is to reformulate the problem as many 2-class problems (one class “versus” all others).
- In certain cases, it is possible that a pattern be assigned to more than one class.

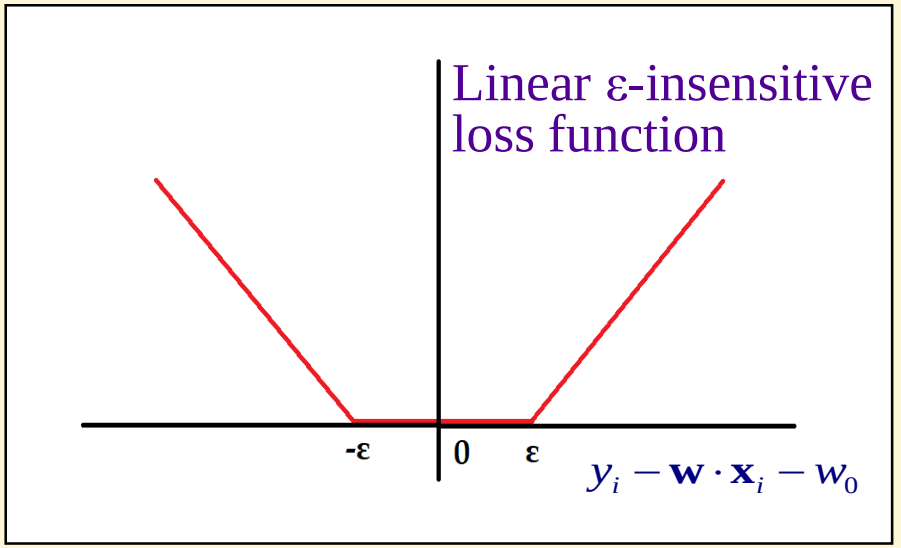
Support vector regression (1)

- In the regression task, we still wish to control model complexity by keeping the norm of the weight vector small enough.
- Additionally, we welcome the existence of patterns with a small vertical deviation from the regressing hyperplane, but we wish to penalize patterns which are far from this hyperplane.
- Thus we need a balance between our desire for minimization of the weight vector and our desire to penalize patterns which are far away from the proposed regressing hyperplane.
- To fulfill our second desire, we employ the so called linear ε-insensitive loss function.

$$L\left(y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0\right) = \max\left(0, \left|y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0\right| - \varepsilon\right)$$

- Patterns corresponding to error less than ε do not get penalized at all.
- Patterns with error greater than ε are penalized in a linear manner with respect to their error (distance from the regressing hyperplane).





Support vector regression (2)

According to our rationale, we wish to minimize the following cost function:

$$\mathsf{E} = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{P} \max\left(0, |y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0| - \varepsilon\right)$$

Regularize

Penalize far away patterns

We introduce auxilliary (slack) variables to facilitate the solution.

- If $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 \geq \varepsilon$ we can write $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 \leq \varepsilon + \xi_i$, with $\xi_i > 0$
- If $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 < \varepsilon$ we can write $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 \leq \varepsilon + \xi_i$, with $\xi_i = 0$
- If $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 \leq -\varepsilon$ we can write $\mathbf{w}\cdot\mathbf{x}_i + w_0 - y_i \leq \varepsilon + \xi_i'$, with $\xi_i' > 0$
- If $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 > -\varepsilon$ we can write $\mathbf{w}\cdot\mathbf{x}_i + w_0 - y_i \leq \varepsilon + \xi_i'$, with $\xi_i' = 0$

The slack variables are zero inside the margin defined by the linear ε-insensitive loss function, and positive outside the margin. They represent the vertical distance from the border of this margin. Therefore we can rewrite the cost function as follows:

$$\mathsf{E} = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{P}\xi_i + \sum_{i=1}^{P}\xi_i'\right)$$

Support vector regression (3)

In all cases, our constraints look like this:

$$\xi_i \geq 0\,,\ y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 \leq \varepsilon + \xi_i$$

$$\xi_i^{'} \geq 0\,,\ \mathbf{w}\cdot\mathbf{x}_i + w_0 - y_i \leq \varepsilon + \xi_i^{'}$$

and the problem becomes:

Minimize:

$$\tfrac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i=1}^{P}\xi_i + \sum_{i=1}^{P}\xi_i^{'}\right)$$

Maximize margin

Penalize far away patterns

Under the constraints: $y_i - \mathbf{w}\cdot\mathbf{x}_i - w_0 \leq \varepsilon + \xi_i$ i.e. $-y_i + \mathbf{w}\cdot\mathbf{x}_i + w_0 + \varepsilon + \xi_i \geq 0$

$\mathbf{w}\cdot\mathbf{x}_i + w_0 - y_i \leq \varepsilon + \xi_i^{'}$ $\quad -\mathbf{w}\cdot\mathbf{x}_i - w_0 + y_i + \varepsilon + \xi_i^{'} \geq 0$

$\xi_i \geq 0 \quad \xi_i^{'} \geq 0$ $\quad \xi_i \geq 0 \quad \xi_i^{'} \geq 0$

Introduce Lagrange multipliers. The corresponding primal Lagrangian is:

$$L(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\xi}^{'}, \boldsymbol{\lambda}, \boldsymbol{\lambda}^{'}, \boldsymbol{\eta}, \boldsymbol{\eta}^{'}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{P}\xi_i + C\sum_{i=1}^{P}\xi_i^{'} - \sum_{i=1}^{P}\eta_i\xi_i - \sum_{i=1}^{P}\eta_i^{'}\xi_i^{'} -$$

$$-\sum_{i=1}^{P}\lambda_i\left(-y_i + \mathbf{w}\cdot\mathbf{x}_i + w_0 + \varepsilon + \xi_i\right) - \sum_{i=1}^{P}\lambda_i^{'}\left(-\mathbf{w}\cdot\mathbf{x}_i - w_0 + y_i + \varepsilon + \xi_i^{'}\right)$$

Support vector regression (4)

The solution is furnished by the conditions:

KARUSH-KUHN-TUCKER CONDITIONS

$$\frac{\partial L}{\partial w_0} = 0$$

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

$$\frac{\partial L}{\partial \xi_i} = 0, \quad i = 1, 2, ..., P$$

$$\frac{\partial L}{\partial \xi_i^{'}} = 0, \quad i = 1, 2, ..., P$$

$$\lambda_i \left(-y_i + \mathbf{w} \cdot \mathbf{x}_i + w_0 + \varepsilon + \xi_i \right) = 0, \quad i = 1, 2, ..., P$$

$$\lambda_i^{'} \left(-\mathbf{w} \cdot \mathbf{x}_i - w_0 + y_i + \varepsilon + \xi_i^{'} \right) = 0, \quad i = 1, 2, ..., P$$

$$\eta_i \xi_i = 0, \quad \eta_i^{'} \xi_i^{'} = 0, \quad i = 1, 2, ..., P$$

$$\lambda_i \geq 0, \quad \lambda_i^{'} \geq 0, \quad \eta_i \geq 0, \quad \eta_i^{'} \geq 0, \quad i = 1, 2, ..., P$$

Support vector regression (5)

$$\frac{\partial L}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^{P} \left(\lambda_i - \lambda_i^{'} \right) = 0 \qquad (A)$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^{P} \left(\lambda_i - \lambda_i^{'} \right) \mathbf{x}_i \qquad (B)$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \lambda_i = C - \eta_i \qquad (C)$$

$$\frac{\partial L}{\partial \xi_i^{'}} = 0 \Rightarrow \lambda_i^{'} = C - \eta_i^{'} \qquad (D)$$

Rewrite Lagrangian to group useful terms together:

$$L = \tfrac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^{P} \left(\lambda_i - \lambda_i^{'} \right) \mathbf{w} \cdot \mathbf{x}_i + \sum_{i=1}^{P} \left(\lambda_i - \lambda_i^{'} \right) y_i - \varepsilon \sum_{i=1}^{P} \left(\lambda_i + \lambda_i^{'} \right) +$$

$$+ \sum_{i=1}^{P} \left(C - \lambda_i - \eta_i \right) \xi_i + \sum_{i=1}^{P} \left(C - \lambda_i^{'} - \eta_i^{'} \right) \xi_i^{'} - w_0 \sum_{i=1}^{P} \left(\lambda_i - \lambda_i^{'} \right)$$

$= 0$ by (C) $= 0$ by (D) $= 0$ by (A)

Support vector regression (6)

Using (B):

$$\tfrac{1}{2}\|\mathbf{w}\|^2 = \tfrac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\left(\lambda_i - \lambda_i^{'}\right)\left(\lambda_j - \lambda_j^{'}\right)(\mathbf{x}_i \cdot \mathbf{x}_j)$$

$$\sum_{i=1}^{P}\left(\lambda_i - \lambda_i^{'}\right)\mathbf{w}\cdot\mathbf{x}_i = \sum_{i=1}^{P}\sum_{j=1}^{P}\left(\lambda_i - \lambda_i^{'}\right)\left(\lambda_j - \lambda_j^{'}\right)(\mathbf{x}_i \cdot \mathbf{x}_j)$$

Substituting, we eliminate all dependence on the primal variables, hence obtaining the dual form of the Lagrangian:

$$L(\mathbf{w}, w_0, \lambda, \lambda^{'}, \eta, \eta^{'}) = -\tfrac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\left(\lambda_i - \lambda_i^{'}\right)\left(\lambda_j - \lambda_j^{'}\right)(\mathbf{x}_i \cdot \mathbf{x}_j) + \sum_{i=1}^{P}\left(\lambda_i - \lambda_i^{'}\right)y_i - \varepsilon\sum_{i=1}^{P}\left(\lambda_i + \lambda_i^{'}\right)$$

Support vector regression (7)

Dual Problem Formulation

Maximize, with respect to $\lambda_i, \lambda_i^{'}$:

$$-\tfrac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\left(\lambda_i-\lambda_i^{'}\right)\left(\lambda_j-\lambda_j^{'}\right)(\mathbf{x}_i\cdot\mathbf{x}_j)+\sum_{i=1}^{P}\left(\lambda_i-\lambda_i^{'}\right)y_i-\varepsilon\sum_{i=1}^{P}\left(\lambda_i+\lambda_i^{'}\right)$$

under the constraints:

$$0\le\lambda_i\le C,\quad 0\le\lambda_i^{'}\le C,\quad i=1,2,...,P$$

$$\sum_{i=1}^{P}\left(\lambda_i-\lambda_i^{'}\right)=0$$

Solution for **w**:

$$\mathbf{w}=\sum_{i=1}^{P}\left(\lambda_i-\lambda_i^{'}\right)\mathbf{x}_i$$

Support vector regression (8)

Prediction for unknown pattern:

$$y_{new} = \mathbf{w} \cdot \mathbf{x}_{new} + w_0 = \sum_{i=1}^{P} \left(\lambda_i - \lambda_i^{'} \right) (\mathbf{x}_i \cdot \mathbf{x}_{new}) + w_0$$

- w_0 is estimated from the patterns that lie on the border of the margin, where the constraints are equal to zero. In practice a w_0 is calculated for each such pattern, and the final w_0 is obtained as the average of all such values.

- Lagrange multipliers for patterns strictly within the margin are zero. This follows from the fact that the $\xi_i, \xi_i^{'}$ are strictly zero within the margin. Therefore the corresponding constraints are not active (they are satisfied as strict inequalities, rather than equalities) and therefore the corresponding Lagrange multipliers are zero by the KKT:

$$\lambda_i \left(-y_i + \mathbf{w} \cdot \mathbf{x}_i + w_0 + \varepsilon + \xi_i \right) = 0, \quad i = 1, 2, ..., P$$

$$\lambda_i^{'} \left(-\mathbf{w} \cdot \mathbf{x}_i - w_0 + y_i + \varepsilon + \xi_i^{'} \right) = 0, \quad i = 1, 2, ..., P$$

- It follows that support vectors are all vectors on the boundary or outside the margin.