Support Vector Machines

IF-3270 Pembelajaran Mesin

Teknik Informatika ITB

Modul 5: Support Vector Machine

Fariska Z. Ruskanda, S.T., M.T.
(fariska@informatika.org)

KK IF -Teknik Informatika - STEI ITB

02 SVM for Linearly Separable Data

IF3270 - Pembelajaran Mesin
(Machine Learning)

Outline

Support Vectors

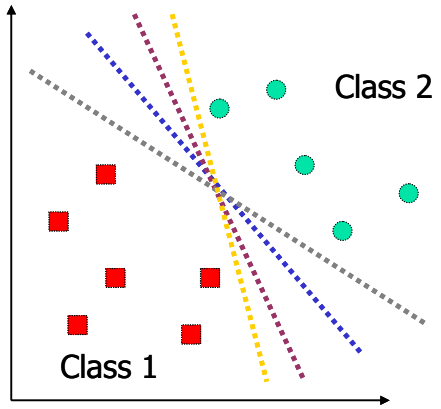
Optimal Hyperplane

Quadratic Optimization Problem

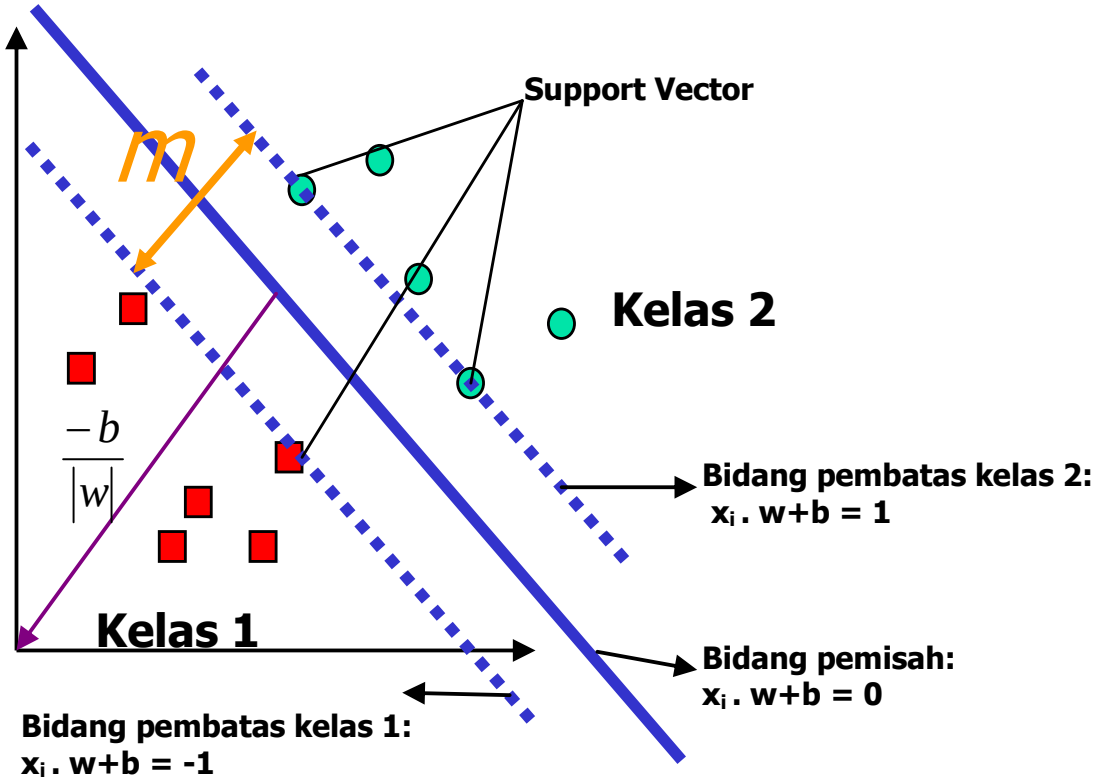
Bidang Pemisah Terbaik

Calculation Example

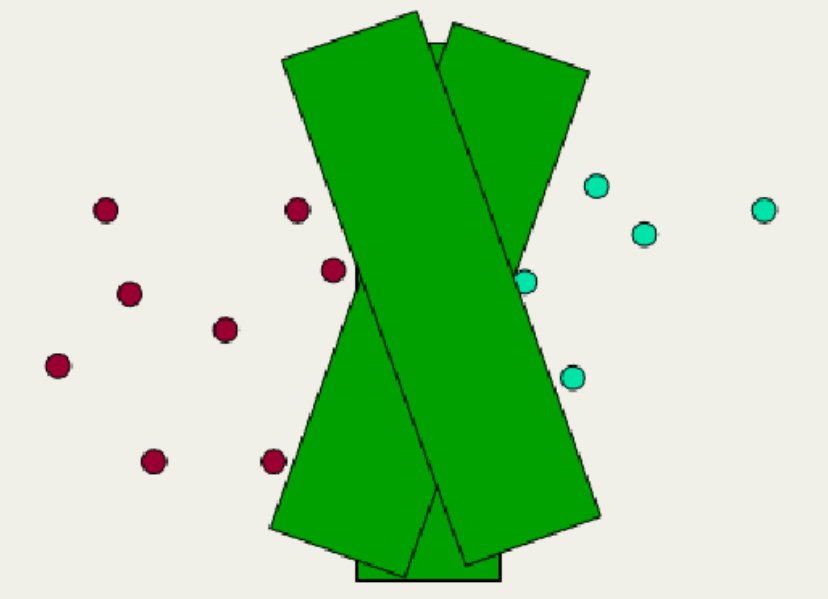
Mencari Bidang Pemisah Terbaik (1)



Mencari bidang pemisah dengan margin (m) terbesar



Mencari Bidang Pemisah Terbaik (2)



- 2 kelas dapat dipisahkan oleh sepasang bidang pembatas yang sejajar.
  - Bidang pembatas pertama membatasi kelas pertama
  - Bidang pembatas kedua membatasi kelas kedua
- ***Support Vector***: Vector pada training data yang men-*support* bidang pemisah

Mengapa mencari margin terbesar?

- Bidang pemisah terbaik dengan margin terbesar memiliki generalisasi yang lebih baik
  - Titik yang dekat dengan bidang pemisah merepresentasikan ketidakyakinan klasifikasi : 50% peluang pengambilan keputusan oleh classifier
- Kapasitas memori untuk menyimpan model menjadi lebih sedikit
  - Hanya support vector yang mempengaruhi pengambilan keputusan

Support Vectors

- The data points that lie closest to the decision surface (or hyperplane)
- They are the data points most difficult to classify
- They have direct bearing on the optimum location of the decision surface

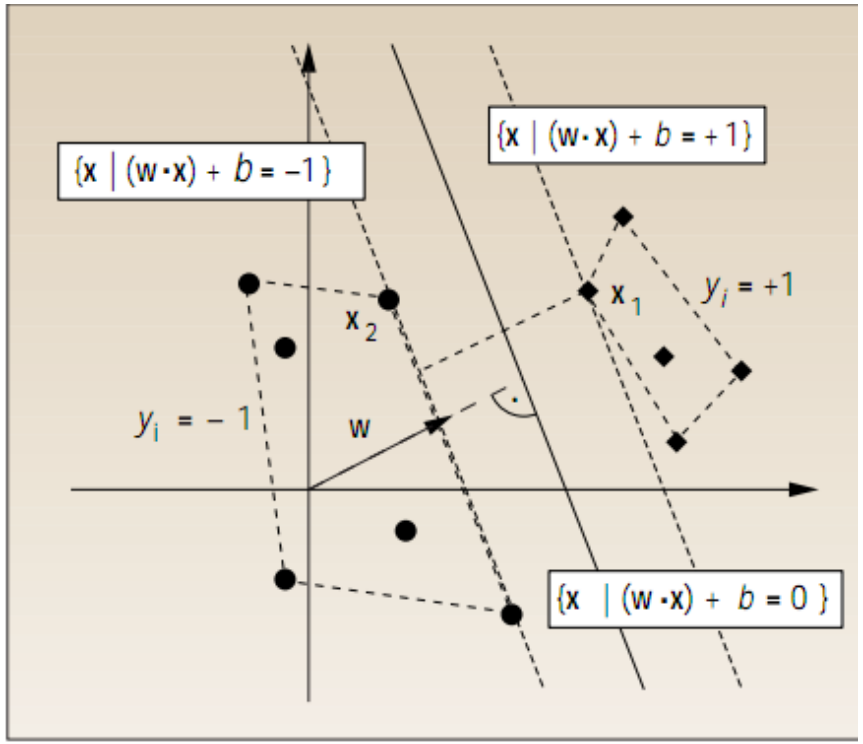
Support Vectors

- Support vectors memenuhi:

$$| \vec{w}.\vec{x}_i + b | = 1$$

- Untuk itu, semua training data berlabel “+” dan “–” memenuhi juga:

$$y_i(\vec{w}.\vec{x}_i + b) \geq 1$$



Optimal Hyperplane

- Optimal hyperplane:
  - *maximum-margin hyperplane*
  - atau hyperplane with maximal margin separation between two classes.
- Maximal margin (2/||w||) ≈ minimize ||w|| dengan batasan konsistensi training data tetap terjaga (semua klasifikasi benar).
- Minimize: $V(\vec{w},b)=\frac{1}{2}\vec{w}.\vec{w}$

  subject to: $y_i(\vec{w}.\vec{x}_i+b)\geq 1$

Recall the distance from a point(x_0,y_0) to a line: $Ax+By+c = 0$ is: $|Ax_0+By_0+c|/\text{sqrt}(A^2+B^2)$, so,
The distance between H_0 and H_1 is then: $|w\bullet x+b|/||w||=1/||w||$, so
The total distance between H_1 and H_2 is thus: $2/||w||$

Optimal Hyperplane (2)

- Pencarian bidang pemisah terbaik dapat dirumuskan menjadi:

$$\min \frac{1}{2}|w|^2$$
$$s.t\ \ y_i(x_i.w+b)-1\geq 0$$

- Want to look for solution point p where
$$\nabla f(p)=\nabla\lambda g(p)$$
$$g(x)=0$$
- Or, combining these two as the *Langrangian* L & requiring derivative of L be zero:
$$L(x,a)=f(x)-ag(x)$$
$$\nabla(x,a)=0$$

- Supaya lebih mudah diselesaikan ubah ke formula Lagrangian

$$\min_{w,b} L_p(w,b,\alpha)\equiv\frac{1}{2}|w|^2-\sum_{i=1}^{n}\alpha_i y_i(x_i.w+b)+\sum_{i=1}^{n}\alpha_i$$
$$\alpha\geq 0$$

- Minimumkan Lp terhadap b dan w (saddle point), diperoleh:

$$\frac{\partial}{\partial b}L_p(w,b,\alpha)=0 \Rightarrow \sum_{i=1}^{n}\alpha_i y_i=0$$

$$\frac{\partial}{\partial w}L_p(w,b,\alpha)=0 \Rightarrow \frac{\partial L_P}{\partial \mathbf{w}}=\mathbf{w}-\sum_{i=1}^{l}a_i y_i\mathbf{x}_i=0$$

$$w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

Langrangian Dual Problem

The Lagrangian Dual Problem: instead of minimizing over **w**, b, subject to constraints involving a's, we can maximize over a (the dual variable) subject to the relations obtained previously for **w** and b

Our solution must satisfy these two relations:

$$\mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{l} a_i y_i = 0$$

By substituting for **w** and b back in the original eqn we can get rid of the dependence on **w** and b.

Note first that we already now have our answer for what the weights **w** must be: they are a linear combination of the training inputs and the training outputs, x_i and y_i and the values of a. We will now solve for the a's by differentiating the dual problem wrt a, and setting it to zero. Most of the a's will turn out to have the value zero. The non-zero a's will correspond to the support vectors

Substitute the Primal Problem

Primal problem:

$$\min L_P = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} a_i y_i \left(\mathbf{x}_i \cdot \mathbf{w} + b\right) + \sum_{i=1}^{l} a_i$$

s.t. $\forall i \; a_i \geq 0$

$$\mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i, \quad \sum_{i=1}^{l} a_i y_i = 0$$

Dual problem:

$$\max L_D(a_i) = \sum_{i=1}^{l} a_i - \frac{1}{2} \sum_{i=1}^{l} a_i a_j y_i y_j \left(\mathbf{x}_i \cdot \mathbf{x}_j\right)$$

s.t. $\sum_{i=1}^{l} a_i y_i = 0 \;\&\; a_i \geq 0$

(note that we have removed the dependence on **w** and b)

Quadratic Optimization Problem

- Minimize: $W(\vec{\alpha}) = -\sum_{i=1}^{n} \alpha_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\vec{x}_i . \vec{x}_j)$

  subject to: $\sum_{i=1}^{n} y_i \alpha_i = 0; \forall i \in [1..n] : \alpha_i \geq 0$

- Support vectors: $\alpha_i > 0$
- Hyperplane: $\vec{w}.\vec{x} = (\sum_{i=1}^{n} \alpha_i y_i \vec{x}_i).\vec{x} = \sum_{i=1}^{n} \alpha_i y_i (\vec{x}_i . \vec{x})$

$$b = y_{sv} - \vec{w}.\vec{x}_{sv}$$

Optimization Problem Example

Suppose we have two 2D data points: [(x11, x12), (x21, x22)] with labels [y1, y2]

- Minimize: W(a) = -(a1+a2) + 0.5*
  [y1*y1*a1*a1*(x11*x11 + x12*x12) +
  y1*y2*a1*a2*(x11*x21 + x12*x22) +
  y2*y1*a2*a1*(x21*x11 + x22*x12) +
  y2*y2*a2*a2*(x21*x21 + x22*x22)]

  subject to: (y1*a1 + y2*a2) = 0
- Support vectors: a1> 0, a2> 0
- Hyperplane:

w.x = a1*y1*((x11,x12)*x) + a2*y2*((x21,x22)*x)

b = ysv – w.xsv

$$W(\vec{\alpha}) = -\sum_{i=1}^{n}\alpha_i + \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j (\vec{x}_i.\vec{x}_j)$$

$$\sum_{i=1}^{n} y_i \alpha_i = 0; \forall i \in [1..n] : \alpha_i \geq 0$$

$$\vec{w}.\vec{x} = (\sum_{i=1}^{n} \alpha_i y_i \vec{x}_i).\vec{x} = \sum_{i=1}^{n} \alpha_i y_i (\vec{x}_i.\vec{x})$$

$$b = y_{sv} - \vec{w}.\vec{x}_{sv}$$

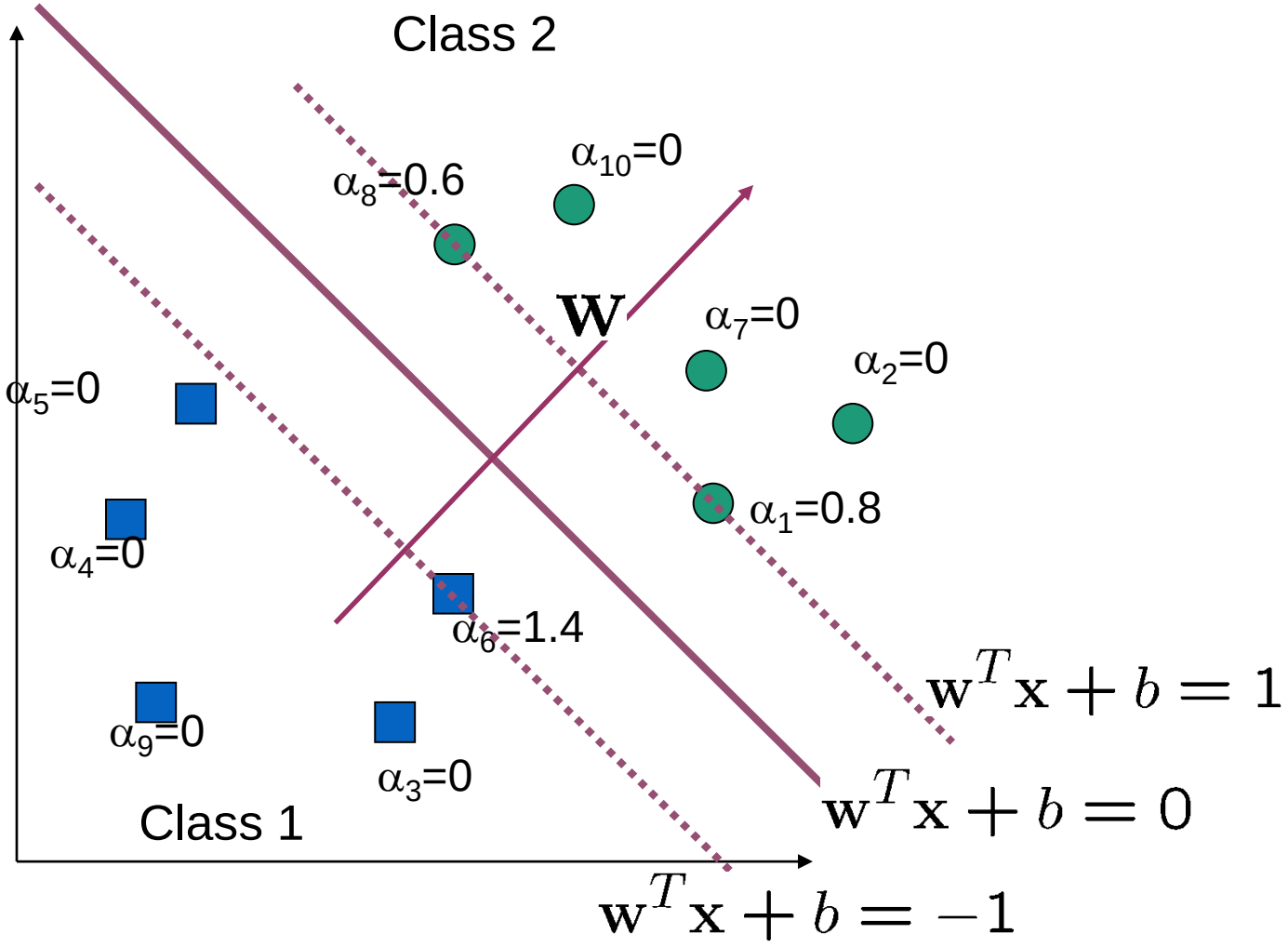
Quadratic Programming

- Quadratic programming (QP) is the problem of optimizing a quadratic objective function and is one of the simplests form of non-linear programming.
- The objective function can contain bilinear or up to second order polynomial terms, and the constraints are linear and can be both equalities and inequalities.

$$\min f(x) = q^T x + \frac{1}{2} x^T Q x$$
$$s.t. Ax = a$$
$$Bx \leq b$$
$$x \geq 0$$

https://optimization.mccormick.northwestern.edu/index.php/Quadratic_programming

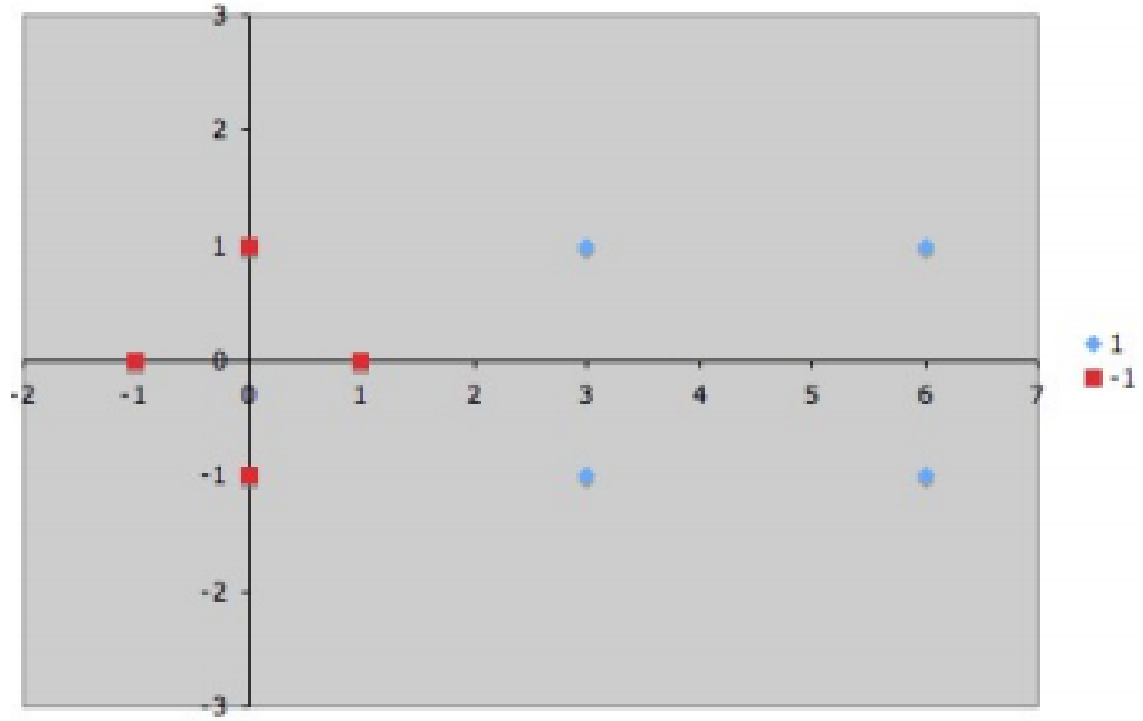
A Geometrical Interpretation



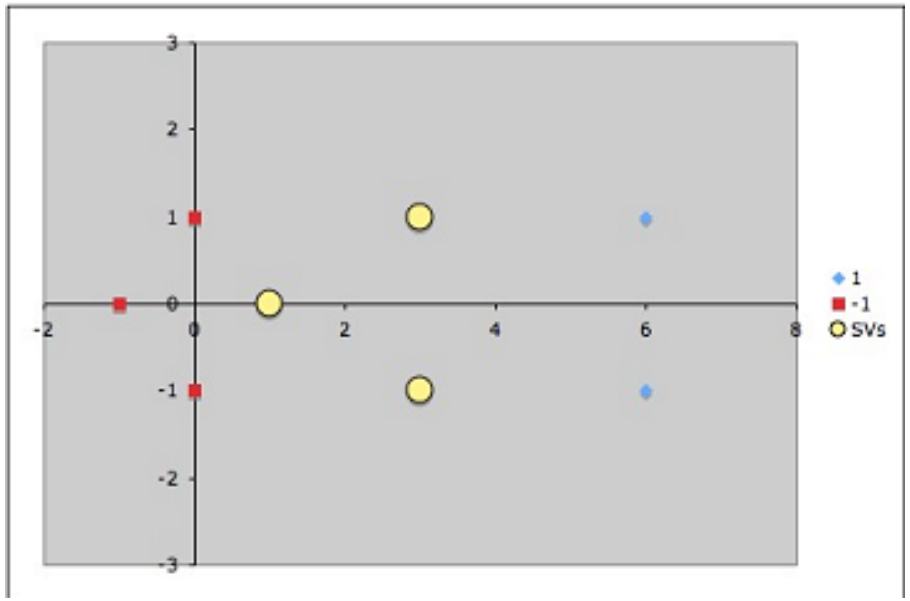
A Simple Introduction to Support Vector Machines; Martin Law; Michigan State Univ.

Contoh

x1	x2	Kelas
3	1	+1
3	-1	+1
6	1	+1
6	-1	+1
1	0	-1
0	1	-1
0	-1	-1
-1	0	-1



- Taken from: http://axon.cs.byu.edu/Dan/678/miscellaneous/SVM.example.pdf



Support Vectors

$$f(\vec{x}) = \sum_{i=1}^{nsv} (\alpha_i y_i \vec{x}_i \vec{x}) + b$$

$$\begin{pmatrix}1\\0\end{pmatrix} \Rightarrow -1 = \alpha_1 . -1 . \begin{pmatrix}1\\0\end{pmatrix} . \begin{pmatrix}1\\0\end{pmatrix} + \alpha_2 . 1 . \begin{pmatrix}3\\1\end{pmatrix} . \begin{pmatrix}1\\0\end{pmatrix} + \alpha_3 . 1 . \begin{pmatrix}3\\-1\end{pmatrix} . \begin{pmatrix}1\\0\end{pmatrix} + b = -\alpha_1 + 3\alpha_2 + 3\alpha_3 + b \ldots (1)$$

$$\begin{pmatrix}3\\1\end{pmatrix} \Rightarrow 1 = \alpha_1 . -1 . \begin{pmatrix}1\\0\end{pmatrix} . \begin{pmatrix}3\\1\end{pmatrix} + \alpha_2 . 1 . \begin{pmatrix}3\\1\end{pmatrix} . \begin{pmatrix}3\\1\end{pmatrix} + \alpha_3 . 1 . \begin{pmatrix}3\\-1\end{pmatrix} . \begin{pmatrix}3\\1\end{pmatrix} + b = -3\alpha_1 + 10\alpha_2 + 8\alpha_3 + b \ldots (2)$$

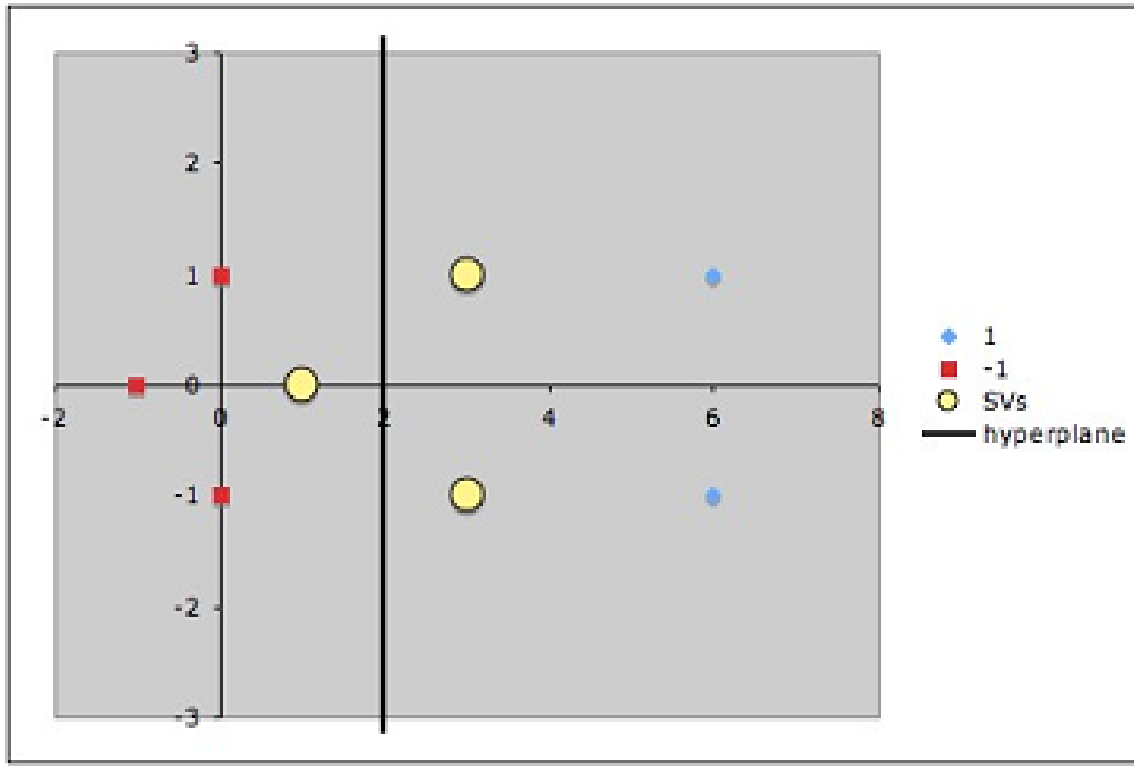
$$\begin{pmatrix}3\\-1\end{pmatrix} \Rightarrow 1 = \alpha_1 . -1 . \begin{pmatrix}1\\0\end{pmatrix} . \begin{pmatrix}3\\-1\end{pmatrix} + \alpha_2 . 1 . \begin{pmatrix}3\\1\end{pmatrix} . \begin{pmatrix}3\\-1\end{pmatrix} + \alpha_3 . 1 . \begin{pmatrix}3\\-1\end{pmatrix} . \begin{pmatrix}3\\-1\end{pmatrix} + b = -3\alpha_1 + 8\alpha_2 + 10\alpha_3 + b \ldots (3)$$

$$-\alpha_1 + 3\alpha_2 + 3\alpha_3 + b = -1 \ldots (1)$$
$$-3\alpha_1 + 10\alpha_2 + 8\alpha_3 + b = 1 \ldots (2)$$
$$-3\alpha_1 + 8\alpha_2 + 10\alpha_3 + b = 1 \ldots (3)$$
$$-\alpha_1 + \alpha_2 + \alpha_3 = 0 \ldots (4)$$

$$(2) - (3): 2\alpha_2 - 2\alpha_3 = 0 \rightarrow \alpha_2 = \alpha_3 \ldots (5)$$
$$(5) subs (1): -\alpha_1 + 6\alpha_2 + b = -1 \ldots (6)$$
$$(5) subs (2): -3\alpha_1 + 18\alpha_2 + b = 1 \ldots (7)$$
$$(6) - (7): 2\alpha_1 - 12\alpha_2 = -2$$
$$\alpha_1 - 6\alpha_2 = -1$$
$$\alpha_1 = 6\alpha_2 - 1 \ldots (8)$$
$$(8,5) subs (4): -6\alpha_2 + 1 + \alpha_2 + \alpha_2 = 0$$
$$-4\alpha_2 = -1 \rightarrow \alpha_2 = \frac{1}{4} = 0.25$$

$$\alpha_1 = 0.5; \alpha_2 = 0.25; \alpha_3 = 0.25$$
$$substitusi (6): -0.5 + 6 * 0.25 + b = -1$$
$$b = -2$$

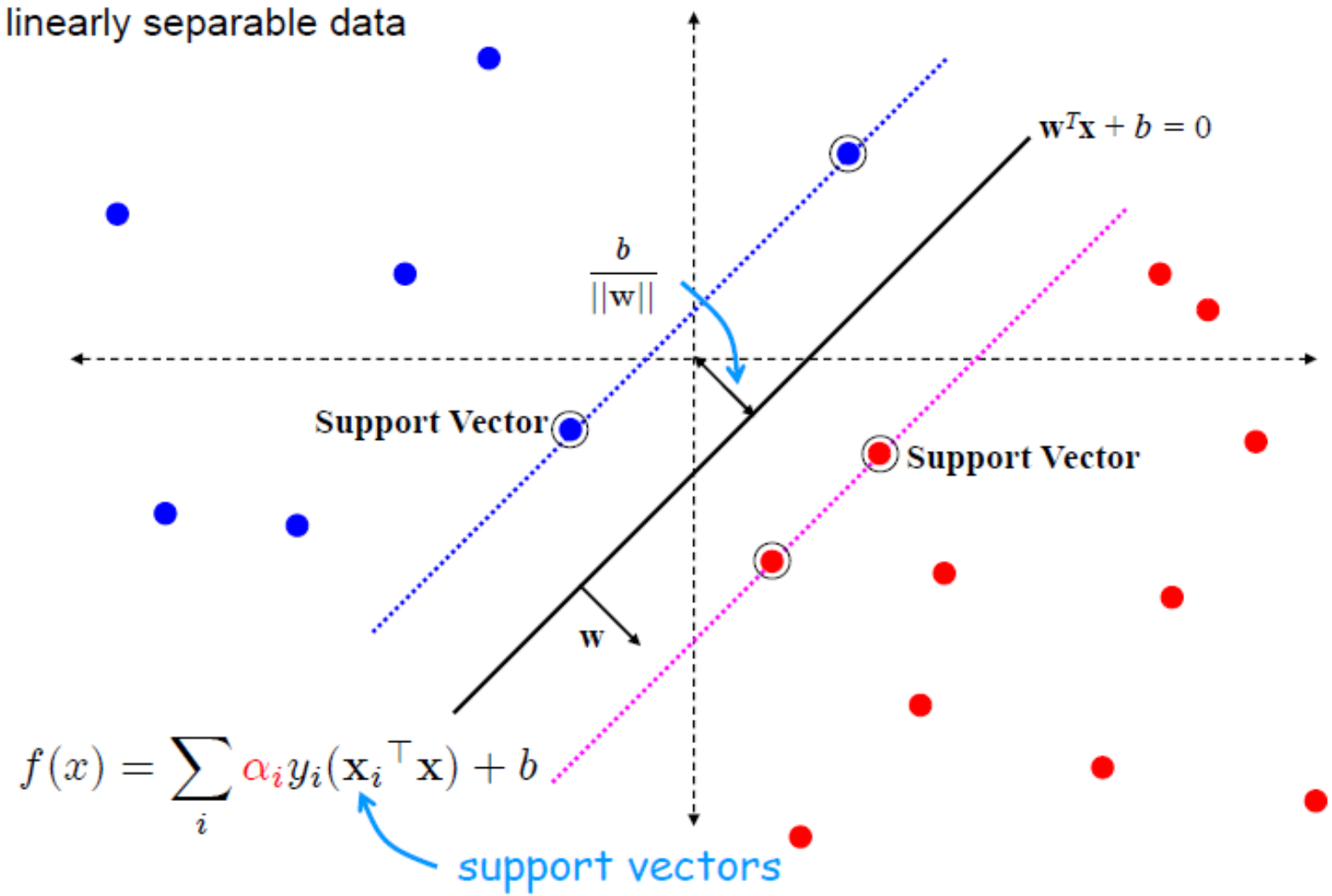
Hipotesis



$$f(\vec{x}) = \sum_{i=1}^{nsv} (\alpha_i y_i \vec{x}_i \vec{x}) - 2; \alpha_1 = 0.5; \alpha_2 = \alpha_3 = 0.25$$

$$f\left(\begin{pmatrix}6\\1\end{pmatrix}\right) = sign((0.5)(-1)\begin{pmatrix}1\\0\end{pmatrix}.\begin{pmatrix}6\\1\end{pmatrix} + (0.25)(1)\begin{pmatrix}3\\1\end{pmatrix}.\begin{pmatrix}6\\1\end{pmatrix} + (0.25)(1)\begin{pmatrix}3\\-1\end{pmatrix}.\begin{pmatrix}6\\1\end{pmatrix} - 2) = sign(-3+4.75+4.25-2) = sign(4) = 1$$

Summary


linearly separable data
$\mathbf{w}^T\mathbf{x} + b = 0$
$\frac{b}{||\mathbf{w}||}$
Support Vector
Support Vector
$\mathbf{w}$
$f(x) = \sum_i \alpha_i y_i (\mathbf{x}_i{}^\top \mathbf{x}) + b$
support vectors

Inner Product and Similiarity

Why should inner product kernels be involved in pattern recognition using SVMs, or at all?

– Intuition is that inner products provide some measure of 'similarity'

– Inner product in 2D between 2 vectors of unit length returns the cosine of the angle between them = how 'far apart' they are

e.g. $\mathbf{x} = [1, 0]^T$, $\mathbf{y} = [0, 1]^T$

i.e. if they are parallel their inner product is 1 (completely similar)

$$\mathbf{x}^T \mathbf{y} = \mathbf{x} \bullet \mathbf{y} = 1$$

If they are perpendicular (completely unlike) their inner product is 0 (so should not contribute to the correct classifier)

$$\mathbf{x}^T \bullet \mathbf{y} = \mathbf{x} \bullet \mathbf{y} = 0$$

03 SVM for Non-linearly Separable Data

IF3270 Pembelajaran Mesin