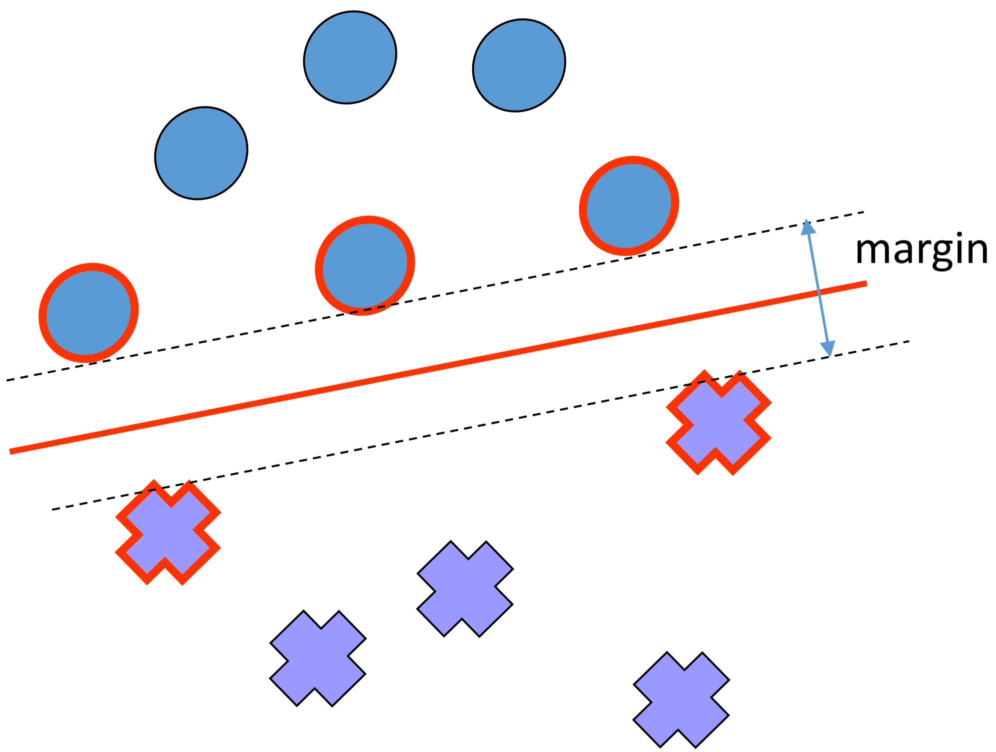
Support Vector Machine
SVM

Masaomi Kimura

SVM（Support vector machines）

- the method to find a border plane to classify two kinds of data



First order equations

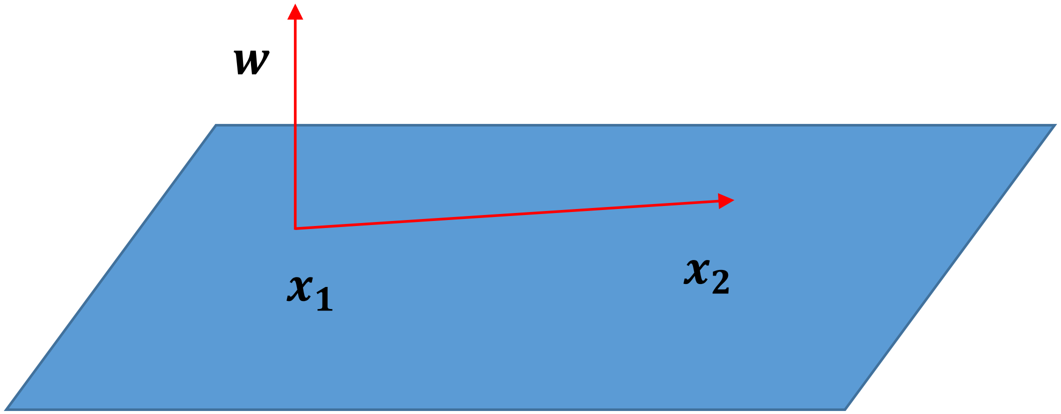
- First order equations express flat things!
  - Line: $2x + 3y = 1$
  - Plane: $2x + 3y + z = 1$
  - Hyper plane: $2x + 3y + z - w = 1$
- Using vectors and their inner product, we can express them easily:

$$\boldsymbol{w} \cdot \boldsymbol{x} - 1 (= 2x + 3y - 1) = 0,$$

where $\boldsymbol{w} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$ and $\boldsymbol{x} = \begin{pmatrix} x \\ y \end{pmatrix}$

Notice

- Vector $\boldsymbol{w}$ is perpendicular to the line/plane $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$



* If $\boldsymbol{w} \cdot \boldsymbol{a} = \boldsymbol{0}$, then $\boldsymbol{w}$ and $\boldsymbol{a}$ are perpendicular to each other.

Problem setting in SVM

- Classify data in two groups in advance, and assign labels $\{y_i | y_i = \pm 1\}$ to each of data
$$\mathbf{x}_\mathrm{i} = \begin{pmatrix} x_1^{(i)} & x_2^{(i)} & \ldots & x_n^{(i)} \end{pmatrix}^T$$
- The data need to be linearly separable
  - SVM can be extended to allow margins or the cases with separable by other shapes

Inequality constraint in SVM

Points having $y_i = +1$ are expressed as $\mathbf{x}_\mathrm{i}^+$

Points having $y_i = -1$ are expressed as $\mathbf{x}_\mathrm{i}^-$

It is possible to find a small positive number ϵ to satisfy

$$\mathbf{w} \cdot \mathbf{x}_\mathrm{i}^+ + b \geq \epsilon \qquad \text{for } y_i = +1$$

$$\mathbf{w} \cdot \mathbf{x}_\mathrm{i}^- + b \leq -\epsilon \qquad \text{for } y_i = -1$$

Because rescaling coefficient does not change the plane, if we rescale $\mathbf{w} \to \epsilon\mathbf{w}$ and $b \to \epsilon b$, we have

$$\mathbf{w} \cdot \mathbf{x}_\mathrm{i}^+ + b \geq 1 \qquad \text{for } y_i = +1$$

$$\mathbf{w} \cdot \mathbf{x}_\mathrm{i}^- + b \leq -1 \qquad \text{for } y_i = -1$$

$$\Rightarrow \quad \boxed{y_i(\mathbf{w} \cdot \mathbf{x}_\mathrm{i} + b) \geq 1}$$

Exercise

- Let us consider the case we have four points and a line:

$$\mathbf{x}_1^+ = (0 \quad 0.5)^T \text{ for } y_1 = +1$$
$$\mathbf{x}_2^+ = (3 \quad 1)^T \text{ for } y_2 = +1$$
$$\mathbf{x}_3^- = (-1 \quad 0.5)^T \text{ for } y_3 = -1$$
$$\mathbf{x}_4^- = (-2 \quad -4)^T \text{ for } y_4 = -1$$
$$2\mathrm{x} + \mathrm{y} + 1 = 0$$

- Let's find a vector $\mathbf{w}$ and a scalar b satisfying

$$y_i(\mathbf{w} \cdot \mathbf{x}_{\mathrm{i}} + b) \geq 1$$

SVM

- Find a hyperplane whose margins μ from both $\mathbf{x}_\mathrm{i}^+$ and $\mathbf{x}_\mathrm{i}^-$ are maximized:

$$\mu = \frac{|\mathbf{w}\cdot\mathbf{x}_\mathrm{i}^+ + b|}{||\mathbf{w}||} + \frac{|\mathbf{w}\cdot\mathbf{x}_\mathrm{i}^- + b|}{||\mathbf{w}||}$$

$$= \frac{\mathbf{w}\cdot\mathbf{x}_\mathrm{i}^+ + b}{||\mathbf{w}||} + \frac{-(\mathbf{w}\cdot\mathbf{x}_\mathrm{i}^- + b)}{||\mathbf{w}||}$$

$$\geq \frac{1}{||\mathbf{w}||} + \frac{-(-1)}{||\mathbf{w}||} = \frac{2}{||\mathbf{w}||} \quad \text{(should be maximized)}$$

Minimize $\frac{1}{2}||\mathbf{w}||^2$ subject to

$$y_i(\mathbf{w}\cdot\mathbf{x}_\mathrm{i} + b) \geq 1$$

Let’s consider the following optimization problem…

- Consider the problem we need to find (x^*, y^*):

$$(x^*, y^*) = \arg\min_{x,y} f(x, y)$$

under the condition

$$g(x, y) \geq 0.$$

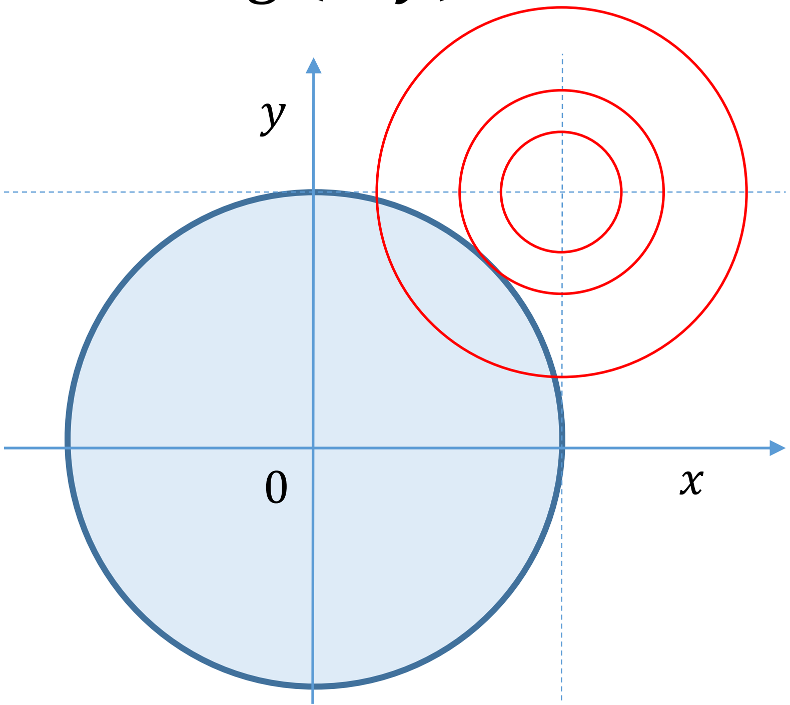
Example

- Consider a problem to minimize

$$f(x, y) = (x-1)^2 + (y-1)^2 = k$$

where x and y satisfy

$$g(x, y) = 1 - x^2 - y^2 \geq 0$$



Which circle has the least k?

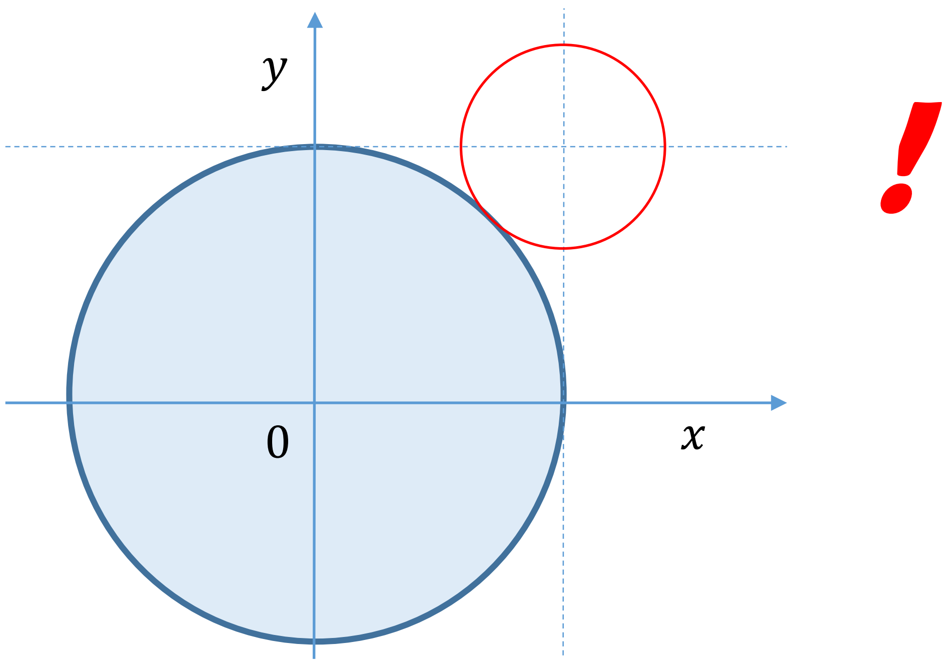
?

Answer

- The circle tangent to the border

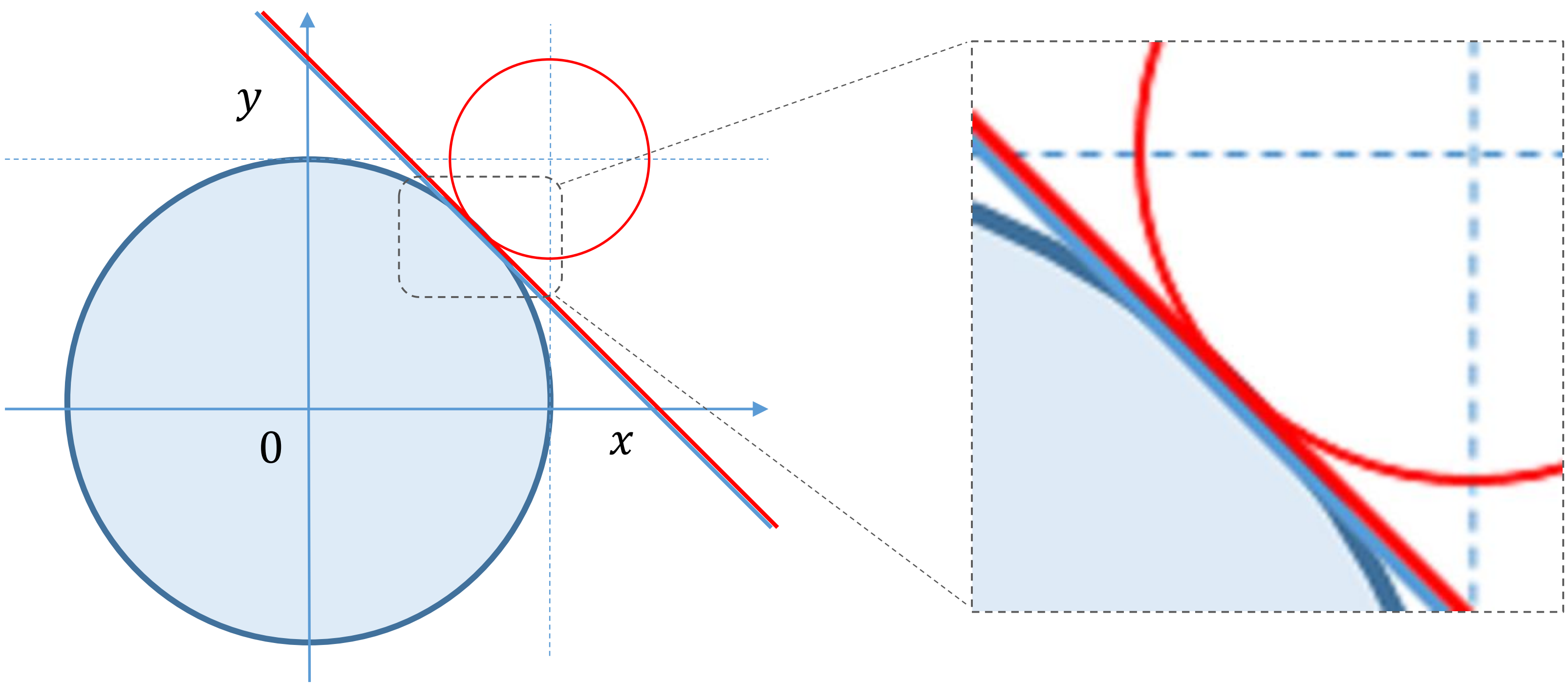
$$g(x, y) = 1 - x^2 - y^2 = 0$$

has the least k



The important thing is...

- The points where two objects (circle-line/circle-circle) were tangent to each other were important.
- At those points, they share the same tangent lines.



Tangent lines

- Around the tangent point (x_0, y_0),
$$g(x_0, y_0) = 0.$$
- These two tangent lines needs to express the same line:
$$\left(\frac{\partial f}{\partial x}(x_0, y_0)\right)(x - x_0) + \left(\frac{\partial f}{\partial y}(x_0, y_0)\right)(y - y_0) = 0$$
$$\left(\frac{\partial g}{\partial x}(x_0, y_0)\right)(x - x_0) + \left(\frac{\partial g}{\partial y}(x_0, y_0)\right)(y - y_0) = 0$$
- Since there is a redundancy of coefficients in an equation to express a line,

$$\frac{\partial f}{\partial x}(x_0, y_0) = \lambda \frac{\partial g}{\partial x}(x_0, y_0) \qquad \frac{\partial f}{\partial y}(x_0, y_0) = \lambda \frac{\partial g}{\partial y}(x_0, y_0)$$

The constraint of λ

- Let the region G^+ be a set of points satisfying the condition $g(x,y) \geq 0$.
- If a point $(x', y') \in G^+$ is close to the tangent point (x_0, y_0),

$$g(x', y') - g(x_0, y_0) \cong \left(\frac{\partial g}{\partial x}(x_0, y_0)\right)(x - x_0) + \left(\frac{\partial g}{\partial y}(x_0, y_0)\right)(y - y_0) \geq 0,$$

  since $g(x', y') \geq 0$ and $g(x_0, y_0)$=0.
- Since $f(x_0, y_0)$ is less than any $f(x', y')$,

$$f(x', y') - f(x_0, y_0) \cong \left(\frac{\partial f}{\partial x}(x_0, y_0)\right)(x - x_0) + \left(\frac{\partial f}{\partial y}(x_0, y_0)\right)(y - y_0) \geq 0.$$

This leads λ to be positive.

In summary…

- To solve a problem to minimize $f(x,y)$ where x and y satisfy $g(x,y) \geq 0$, we should solve

$$\frac{\partial f}{\partial x}(x,y) - \lambda \frac{\partial g}{\partial x}(x,y) = 0$$

$$\frac{\partial f}{\partial y}(x,y) - \lambda \frac{\partial g}{\partial y}(x,y) = 0$$

$$\lambda > 0$$

$$g(x,y) = 0.$$

- Notice that the point (x,y) minimizing $f(x,y)$ without conditions sits outside G^+.

Another case we need to consider

- Consider the case that the point (x, y) minimizing $f(x, y)$ without conditions sits inside G^+.
- It's super easy!! We can use a USUAL differential technique.

$$
\begin{aligned}
&\frac{\partial f}{\partial x}(x, y) = 0 \\
&\frac{\partial f}{\partial y}(x, y) = 0 \\
&g(x, y) > 0
\end{aligned}
\quad \Longrightarrow \quad
\begin{aligned}
&\frac{\partial f}{\partial x}(x, y) - \lambda \frac{\partial g}{\partial x}(x, y) = 0 \\
&\frac{\partial f}{\partial y}(x, y) - \lambda \frac{\partial g}{\partial y}(x, y) = 0 \\
&\lambda = 0 \\
&g(x, y) > 0
\end{aligned}
$$

Kuhn-Tucker conditions

- In the case we need to find (x^*, y^*):
$$(x^*, y^*) = \arg \min_{x,y} f(x, y)$$
under the condition $g(x, y) \geq 0$, we find it by solving

$$\frac{\partial f}{\partial x}(x^*, y^*) - \lambda \frac{\partial g}{\partial x}(x^*, y^*) = 0$$

$$\frac{\partial f}{\partial y}(x^*, y^*) - \lambda \frac{\partial g}{\partial y}(x^*, y^*) = 0$$

$$\lambda \geq 0$$

$$\lambda \cdot g(x^*, y^*) = 0.$$

Or…

- Define a function $L(x, y, \lambda) = f(x, y) - \lambda g(x, y)$, then the equations to be solved are:

$$\frac{\partial L}{\partial x}(x^*, y^*, \lambda^*) \ = 0$$

$$\frac{\partial L}{\partial y}(x^*, y^*, \lambda^*) \ = 0$$

$$\lambda^* \cdot \frac{\partial L}{\partial \lambda}(x^*, y^*, \lambda^*) = 0$$

$$\lambda^* \geq 0$$

More generally…

- Using a vector $\mathbf{x}$, define a function
$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_i \lambda_i g_i(\mathbf{x}),$$
then the equations to be solved are:

$$\frac{\partial L}{\partial \mathbf{x}}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$$

$$\lambda_i^* \cdot \frac{\partial L}{\partial \lambda_i}(\mathbf{x}^*, \boldsymbol{\lambda}^*) = 0$$

$$\lambda_i^* \geq 0$$

Then we can extend the discussion to a space whose dimension is more than two.

Dual theory

- To solve it,

$$\frac{\partial L}{\partial \mathbf{x}}(\mathbf{x}, \boldsymbol{\lambda}) = 0 \Rightarrow \mathbf{x} = \boldsymbol{h}(\boldsymbol{\lambda})$$

- If we consider the case $\lambda_i > 0$, we solve $g_i(\boldsymbol{h}(\boldsymbol{\lambda})) = 0$ to find $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$.
- Dual function is defined as

$$\theta(\boldsymbol{\lambda}) = L(\boldsymbol{h}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = f(\boldsymbol{h}(\boldsymbol{\lambda})) - \sum_i \lambda_i g_i(\boldsymbol{h}(\boldsymbol{\lambda}))$$

$$\frac{\partial \theta(\boldsymbol{\lambda})}{\partial \lambda_i} = \left.\frac{\partial L}{\partial \mathbf{x}}\right|_{\mathbf{x}=\boldsymbol{h}} \frac{d\boldsymbol{h}}{d\lambda_i} - g_i\big(\boldsymbol{h}(\boldsymbol{\lambda})\big) = -g_i\big(\boldsymbol{h}(\boldsymbol{\lambda})\big) \leq \mathbf{0}$$

Dual theory (cont')

- Let $\boldsymbol{\lambda}^*$ be a solution of $\frac{\partial\theta(\boldsymbol{\lambda})}{\partial\boldsymbol{\lambda}} = 0$. Then $\theta(\boldsymbol{\lambda}^*)$ takes a maximal value:
  - $\theta(\boldsymbol{\lambda}^*) = \theta(\boldsymbol{\lambda}) + \frac{\partial\theta(\boldsymbol{\lambda})}{\partial\lambda_i}(\lambda_i^* - \lambda_i) \geq \theta(\boldsymbol{\lambda})$
- Since $f(\boldsymbol{h}(\boldsymbol{\lambda})) \geq f(\boldsymbol{h}(\boldsymbol{\lambda})) - \sum_i \lambda_i g_i(\boldsymbol{h}(\boldsymbol{\lambda})) = \theta(\boldsymbol{\lambda})$, $f\big(\boldsymbol{h}(\boldsymbol{\lambda}^*)\big) = \theta(\boldsymbol{\lambda}^*)$. The $f\big(\boldsymbol{h}(\boldsymbol{\lambda}^*)\big)$ takes a minimum value, because

$$\left.\frac{\partial L}{\partial \mathbf{x}}\right|_{\mathbf{x}=\boldsymbol{h}(\boldsymbol{\lambda}^*)} = \mathbf{0} \quad \text{and} \quad \left.\frac{\partial\theta(\boldsymbol{\lambda})}{\partial\lambda_i}\right|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}^*} = -g_i\big(\boldsymbol{h}(\boldsymbol{\lambda}^*)\big) = \mathbf{0}$$

Differential by a vector

- Let a vector denote $\mathbf{x} = (x_1 \quad x_2 \quad \cdots \quad x_n)^T$
  - $\mathbf{x}^T$ is a transposed vector (matrix) of $\mathbf{x}$.
- Definition:

$$\frac{\partial f}{\partial \mathbf{x}} = \left(\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n} \right)^T$$

- Examples

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{w} \cdot \mathbf{w}) = \left(\frac{\partial}{\partial w_1} \sum_i w_i^2 \quad \frac{\partial}{\partial w_2} \sum_i w_i^2 \quad \cdots \quad \frac{\partial}{\partial w_n} \sum_i w_i^2 \right)^T$$

$$= (2w_1 \quad 2w_2 \quad \cdots \quad 2w_n)^T = 2\mathbf{w}$$

$$\frac{\partial}{\partial \mathbf{w}}(\mathbf{x} \cdot \mathbf{w}) = \mathbf{x}$$

Application of Kuhn-Tucker cond.

Define $L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_i \alpha_i\{y_i(\mathbf{w}\cdot\mathbf{x}_\mathrm{i} + b) - 1\}$, then KT condition is:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i y_i \mathbf{x}_\mathrm{i} = 0$$

$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$

$$\alpha_i\{y_i(\mathbf{w}\cdot\mathbf{x}_\mathrm{i} + b) - 1\} = 0$$

Dual expression of SVM

- Inserting $\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_\mathrm{i}$ into L, we get a dual expression based on Kuhn-Tucker condition, which should be maximized is:

$$\theta(\boldsymbol{\alpha}) = \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i y_i \alpha_j y_j (\mathbf{x}_\mathrm{i} \cdot \mathbf{x}_\mathrm{j})$$

$$\alpha_i \geq 0$$

$$\alpha_i\{y_i(\mathbf{w} \cdot \mathbf{x}_\mathrm{i} + b) - 1\} = 0$$

$$\sum_i \alpha_i y_i = 0$$

$\alpha_i \neq 0$ requires $\mathbf{x}_\mathrm{i}$ to satisfy

$$y_i(\mathbf{w} \cdot \mathbf{x}_\mathrm{i} + b) - 1 = 0$$

(support vectors)

After obtaining a solution

- We can calculate coefficients of the hyperplane:

$$\mathbf{w} = \sum_i \alpha_i y_i \mathbf{x}_\mathrm{i}$$

$$b = -\frac{\min(\mathbf{w} \cdot \mathbf{x}_\mathrm{i}^{+}) + \max(\mathbf{w} \cdot \mathbf{x}_\mathrm{i}^{-})}{2}$$

- A class for new data $\mathbf{x}$ is given by the value

$$H(x) = sign\left(\sum_i \alpha_i y_i (\mathbf{x}_\mathrm{i} \cdot \mathbf{x}) + b\right)$$

After obtaining a solution

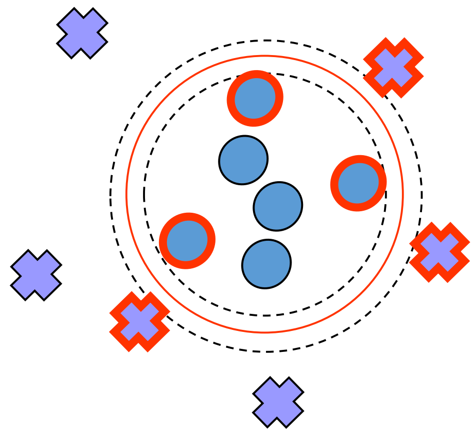
- We can calculate coefficients of the hyperplane and get $\mathbf{w}$ and b.

- A class for new data $\mathbf{x}$ is given by the value

$$H(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

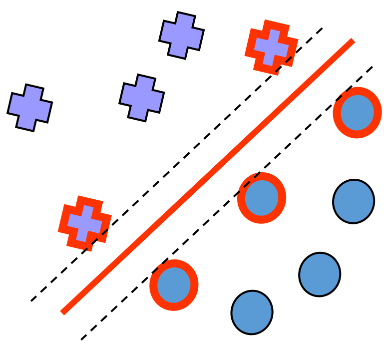
$$sign(x) = \begin{cases} 1 & (x \geq 0) \\ -1 & (x < 0) \end{cases}$$

Extension to a high dimensional space



$$\vec{x} = (x, y)$$

$$(x-a)^2 + (y-b)^2 = r^2$$



$$\phi(x) = (X, Y, Z, V)$$

$$X + Y + aZ + bV + c = 0$$

$$(X = x^2, Y = y^2, Z = -2x, V = -2y, c = a^2 + b^2 - r^2)$$

Generalization

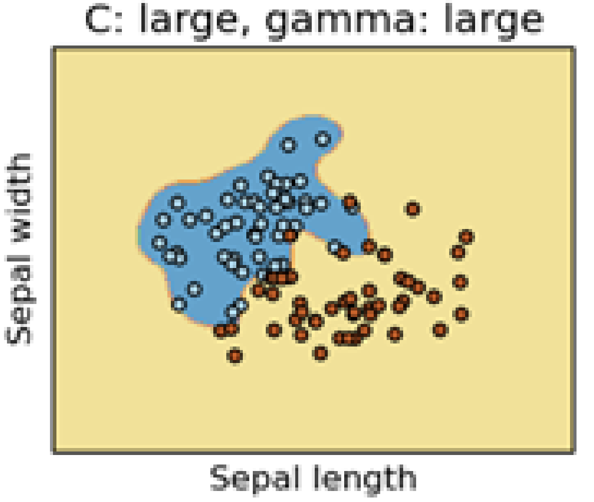
- If we can map $\mathbf{x}$ to a higher dimensional space
$$\mathbf{X} = \boldsymbol{\phi}(\mathbf{x})$$
and data in two classes are separable by a hyper plane, the problem is to find the hyper plane:
$$\boldsymbol{W} \cdot \boldsymbol{\phi}(\mathbf{x}) + b = 0,$$
which maximizes margin between the two classes.
- By the detail discussion (see Appendix), $\boldsymbol{W}$ is given as $\mathbf{W} = \sum_i \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_\mathrm{i})$.

$$H(\mathbf{x}) = sign\left(\sum_i \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_\mathrm{i}) \cdot \boldsymbol{\phi}(\mathbf{x}) + b\right)$$

RBF kernel

- The properties of a kernel $K(\mathbf{x}, \mathbf{y}) = \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{y})$ needs to have are:
  - $K(\mathbf{x}, \mathbf{x}) \geq 0$
  - $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$
- The most used kernel is RBF (radial based function) kernel defined as

  $K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{y}\|^2)$



https://qiita.com/sz_dr/items/f3d6630137b184156a67