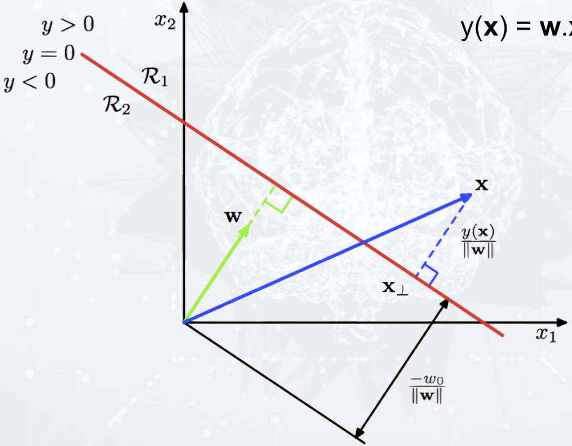
Machine Learning

Support Vector Machine

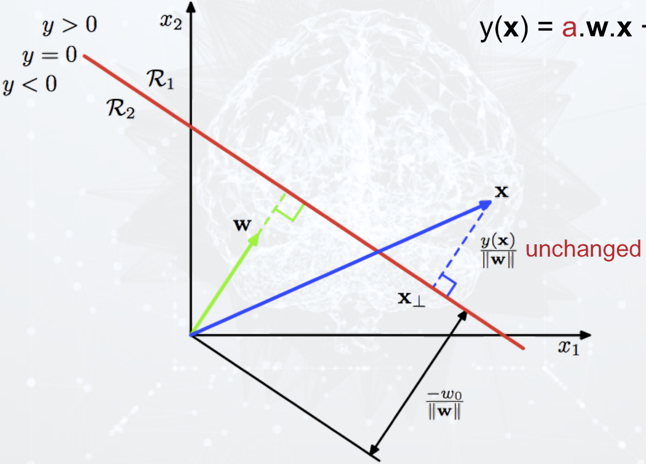
Lecturer: Duc Dung Nguyen, PhD.
Contact: nddung@hcmut.edu.vn

Faculty of Computer Science and Engineering
Hochiminh city University of Technology

Analytical Geometry



$y(\mathbf{x}) = \mathbf{w}.\mathbf{x} + w_0$



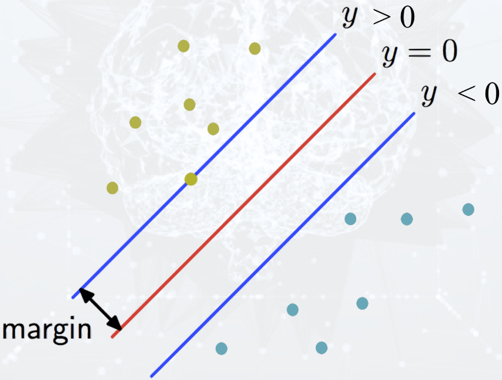
$y(\mathbf{x}) = a.\mathbf{w}.\mathbf{x} + a.w_0$

Maximum Margin Classifiers

- Assume that the data are **linearly separable**
- **Decision boundary** equation:

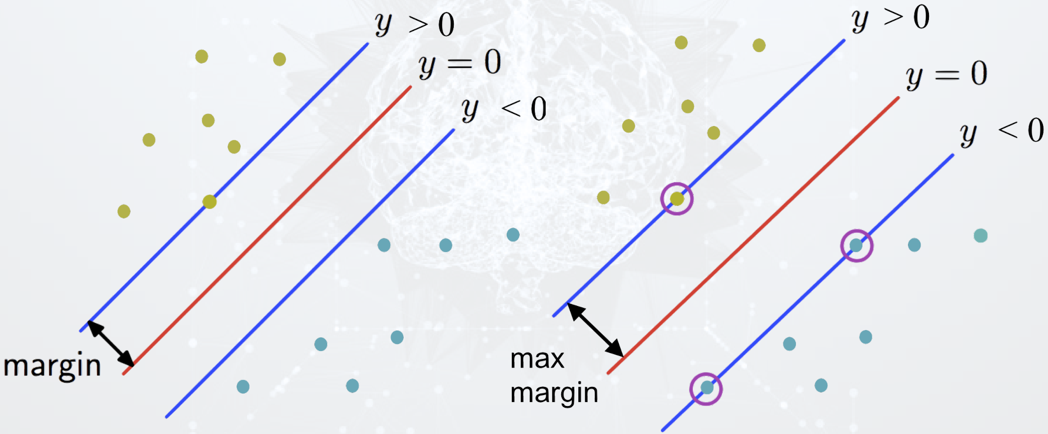
$$y(x) = w.x + b$$

- **Margin**: the smallest distance between the decision boundary and any of the samples.

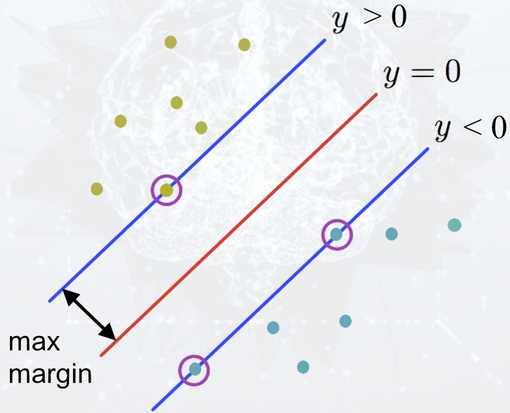


Maximum margin classifiers

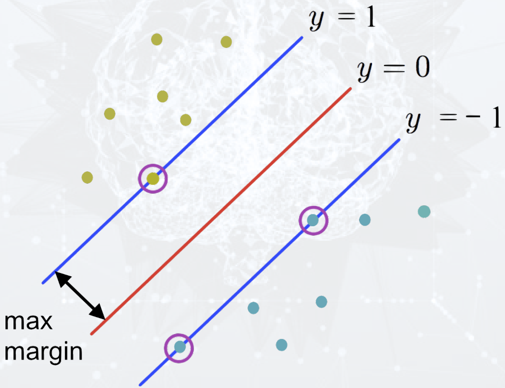
- **Margin**: the smallest distance between the decision boundary and any of the samples.



- **Support vectors**: samples at the two margins.



- *Scaling* **y (support vectors)** to be 1 or -1:



- **Signed distance** between the decision boundary and a sample x_n:

$$\frac{y(x_n)}{||w||}$$

- **Signed distance** between the decision boundary and a sample x_n:

$$\frac{y(x_n)}{||w||}$$

- **Absolute distance** between the decision boundary and a sample x_n:

$$\frac{t_n.y(x_n)}{||w||}$$

$t_n = +1$ iff $y(x_n) > 0$ **and** $t_n = -1$ iff $y(x_n) < 0$

- Maximum margin:

$$\arg\max_{w,b} \left\{ \frac{1}{||w||} min_n (t_n.(w.x_n + b)) \right\}$$

with the constraint:

$$t_n.(w.x_n + b) \geq 1$$

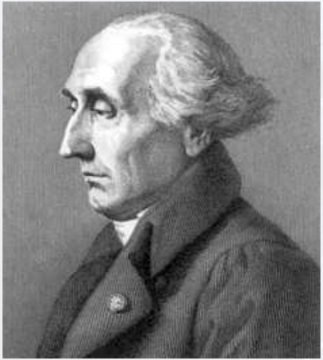
- To be optimized:

$$\arg\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2$$

with the constraint:

$$t_n.(\mathbf{w}.\mathbf{x}_n + b) \geq 1$$

Lagrange Multipliers



Joseph-Louis Lagrange born 25 January 1736 – Paris, 10 April 1813; also reported as Giuseppe Luigi Lagrange, was an Italian Enlightenment Era mathematician and astronomer. He made significant contributions to the fields of analysis, number theory, and both classical and celestial mechanics.

- Problem:

$$\arg\max_x f(x)$$

with the constraint:

$$g(x) = 0$$

- Solution is the stationary point of the Lagrange function:

$$L(x, \lambda) = f(x) + \lambda.g(x)$$

such that:

$$\partial L(x, \lambda)/\partial x_n = \partial f(x)/\partial x_n + \lambda.\partial g(x)/\partial x_n = 0$$

and

$$\partial L(x, \lambda)/\partial \lambda = g(x) = 0$$

Optimization using Lagrange multipliers

- Example:

$$f(x) = 1 - u^2 - v^2$$

with the constraint:

$$g(x) = u + v - 1 = 0$$

- Lagrange function:

$$
\begin{aligned}
L(x,\lambda) &= f(x) + \lambda.g(x) = (1 - u^2 - v^2) + \lambda.(u + v - 1)\\
\partial L(x,\lambda)/\partial u &= \partial f(x)/\partial u + \lambda.\partial g(x)/\partial u = -2u + \lambda = 0\\
\partial L(x,\lambda)/\partial v &= \partial f(x)/\partial v + \lambda.\partial g(x)/\partial v = -2v + \lambda = 0\\
\partial L(x,\lambda)/\partial \lambda &= g(x) = u + v - 1 = 0
\end{aligned}
$$

- Solution: $u = 1/2$ and $v = 1/2$

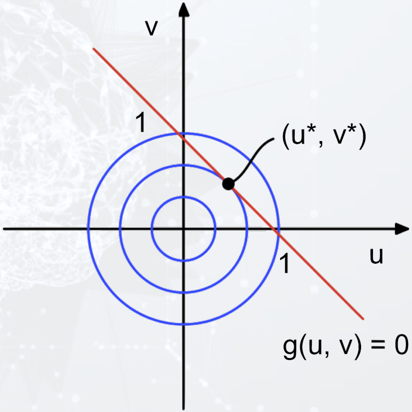
Optimization using Lagrange multipliers

- Example:

$$f(x) = 1 - u^2 - v^2$$

with the constraint:

$$g(x) = u + v - 1 = 0$$



- Problem:

$$\arg\max_x f(x)$$

  with the **inequality constraint**:

$$g(x) \geq 0$$

Optimization using Lagrange multipliers

Solution is the stationary point of the Lagrange function:

$$L(x, \lambda) = f(x) + \lambda.g(x)$$

such that:

$$\partial L(x, \lambda)/\partial x_n = \partial f(x)/\partial x_n + \lambda.\partial g(x)/\partial x_n = 0$$

and

$$g(x) \geq 0$$

$$\lambda \geq 0$$

$$\lambda.g(x) = 0$$

Optimization using Lagrange multipliers

- To be optimized:

$$\arg\min_{w,b} \frac{1}{2}\|\mathbf{w}\|^2$$

with the constraint:

$$t_n.(\mathbf{w}.\mathbf{x}_n + b) \geq 1)$$

- Lagrange function for maximum margin classifier:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1..N} a_n.(t_n.(\mathbf{w}.\mathbf{x}_n + b) - 1)$$

$$t_n.(\mathbf{w}.\mathbf{x}_n + b) - 1 \geq 0$$

$$a_n \geq 0$$

$$a_n.(t_n.(\mathbf{w}.\mathbf{x}_n + b) - 1) = 0$$

- Lagrange function for maximum margin classifier:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1..N} a_n.(t_n.(\mathbf{w}.x_n + b) - 1)$$

- Solution for w:

$$\partial(\mathbf{w}, b, \mathbf{a})/\partial\mathbf{w} = 0$$

$$\mathbf{w} = \sum_{n=1..N} a_n.t_n.x_n$$

$$\partial L(\mathbf{w}, b, \mathbf{a})/\partial b = \sum_{n=1..N} a_n.t_n = 0$$

Optimization using Lagrange multipliers

- Lagrange function for maximum margin classifier:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1..N} a_n.(t_n.(\mathbf{w}.x_n + b) - 1)$$

- Solution for a: dual representation to be optimized

$$L^*(\mathbf{a}) = \sum_{n=1..N} a_n - \frac{1}{2} \sum_{n=1..N} \sum_{m=1..N} a_n.a_m.t_n.t_m.x_n.x_m$$

with the constraints:

$$a_n \geq 0$$

$$\sum_{n=1..N} a_n.t_n = 0$$

Optimization using Lagrange multipliers

- Lagrange function for maximum margin classifier:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1..N} a_n.(t_n.(\mathbf{w}.x_n + b) - 1)$$

- Solution for a: dual representation to be optimized

$$L^*(\mathbf{a}) = \sum_{n=1..N} a_n - \frac{1}{2} \sum_{n=1..N} \sum_{m=1..N} a_n.a_m.t_n.t_m.x_n.x_m$$

  Why optimization via dual representation?

- Sparsity: $a_n = 0$ if x_n is not a support vector.

- Lagrange function for maximum margin classifier:

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{n=1..N} a_n.(t_n.(\mathbf{w}.x_n + b) - 1)$$

$$a_n.(t_n.(\mathbf{w}.x_n + b) - 1) = 0$$

- Solution for b:

$$b = \frac{1}{|S|}\sum_{n \in S} a_m.t_m.x_m.x_n$$

where S is the set of support vectors ($a_n \neq 0$)

- Classification:

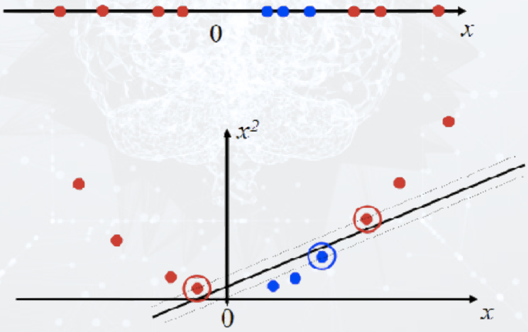
$$y(x) = \mathbf{w}.\mathbf{x} + b = \sum_{n=1..N} a_n.t_n.x_n.x + b$$

$$y(x) > 0 \rightarrow +1$$

$$y(x) < 0 \rightarrow -1$$

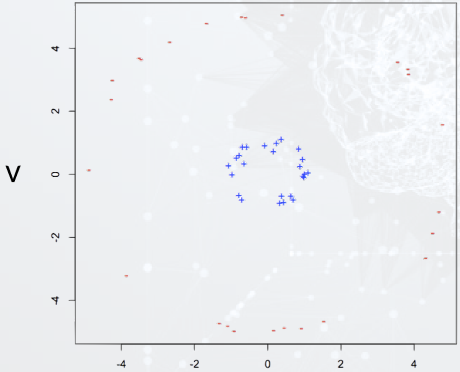
Non-linearly Separable Data

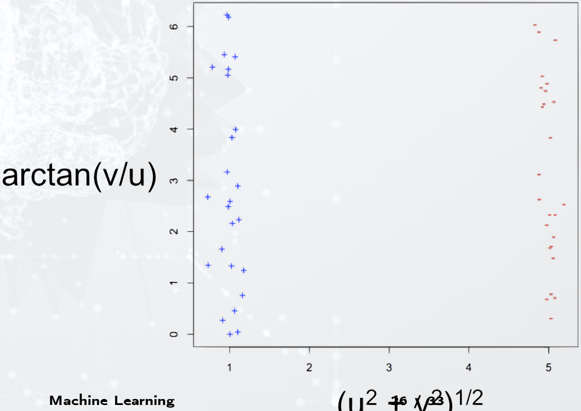
- Mapping the data points into a **high dimensional** feature space.
- Example 1:
  - Original space: (x)
  - New space: (x, x^2)



Kernel trick for non-linearly separable data

- Example 2:
  - Original space: (u, v)
  - New space: $((u^2 + v^2)^{1/2}, \arctan(v/u))$



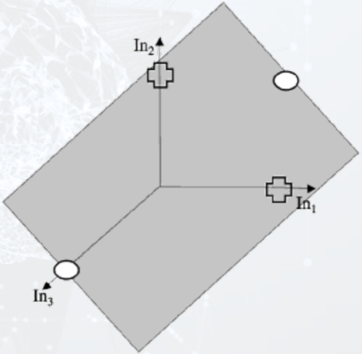


Example 3: **XOR** function

In1	In2	t
0	0	0
0	1	1
1	0	1
1	1	0

Example 3: **XOR** function

In1	In2	In3	Output
0	0	1	1
0	1	0	0
1	0	0	0
1	1	0	1



Kernel trick for non-linearly separable data

- Classification in the new space:

$$y(x) = w.\phi(x) + b = \sum_{n=1..N} a_n.t_n.\phi(x_n).\phi(x) + b$$

Kernel trick for non-linearly separable data

- Classification in the new space:

$$y(x) = w.\phi(x) + b = \sum_{n=1..N} a_n.t_n.\phi(x_n).\phi(x) + b$$

- Computational complexity of $\phi(x_n).\phi(x)$ is high due to the high dimension of $\phi(.)$.

Kernel trick for non-linearly separable data

- Classification in the new space:

$$y(x) = w.\phi(x) + b = \sum_{n=1..N} a_n.t_n.\phi(x_n).\phi(x) + b$$

- Computational complexity of $\phi(x_n).\phi(x)$ is high due to the high dimension of $\phi(.)$.
- Kernel trick:

$$\phi(x_n).\phi(x_m) = K(x_n, x_m)$$

Kernel trick for non-linearly separable data

- A typical kernel function:

$$K(u,v) = (1 + u.v)^2$$

$$\begin{aligned}\phi((u_1.u_2,...,u_d)) = (1, \sqrt{2}u_1, \sqrt{2}u_2, ..., \sqrt{2}u_d,\\ \sqrt{2}u_1.u_2, \sqrt{2}u_1.u_3, ..., \sqrt{2}u_{d-1}.u_d,\\ u_1^2, u_2^2, ..., u_d^2)\end{aligned}$$

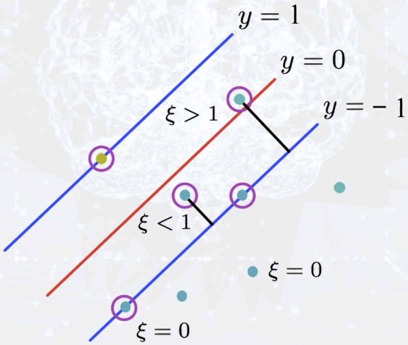
$$\phi(u).\phi(v) = 1 + 2\sum_{i=1..d} u_i.v_i + 2\sum_{i=1..d-1}\sum_{j=i+1..d} u_i.v_i.u_j.v_j + \sum_{i=1..d} u_i^2 v_i^2$$

$$\phi(u.\phi(v) = K(u,v)$$

- Is $\phi(x)$ guaranteed to be separable?

Soft-margin

- Soft-margin SVM: to allow some of the training samples to be misclassified.
- Slack variable: ξ



Soft margin SVM

- New constraints:

$$t_n.(w.x_n + b) \geq 1 - \xi_n$$

$$\xi_n \geq 0$$

Soft margin SVM

- New constraints:

$$t_n.(w.x_n + b) \geq 1 - \xi_n$$

$$\xi_n \geq 0$$

- To be minimized:

$$\frac{1}{2}||w||^2 = C \sum_{n=1..N} \xi_n$$

$C > 0$: controls the trade-off between the margin and slack variable penalty

Summary

- SVM is a sparse kernel method.
- Soft margin SVM is to deal with non-linearly separable data after kernel mapping.