SUPPORT VECTOR MACHINES

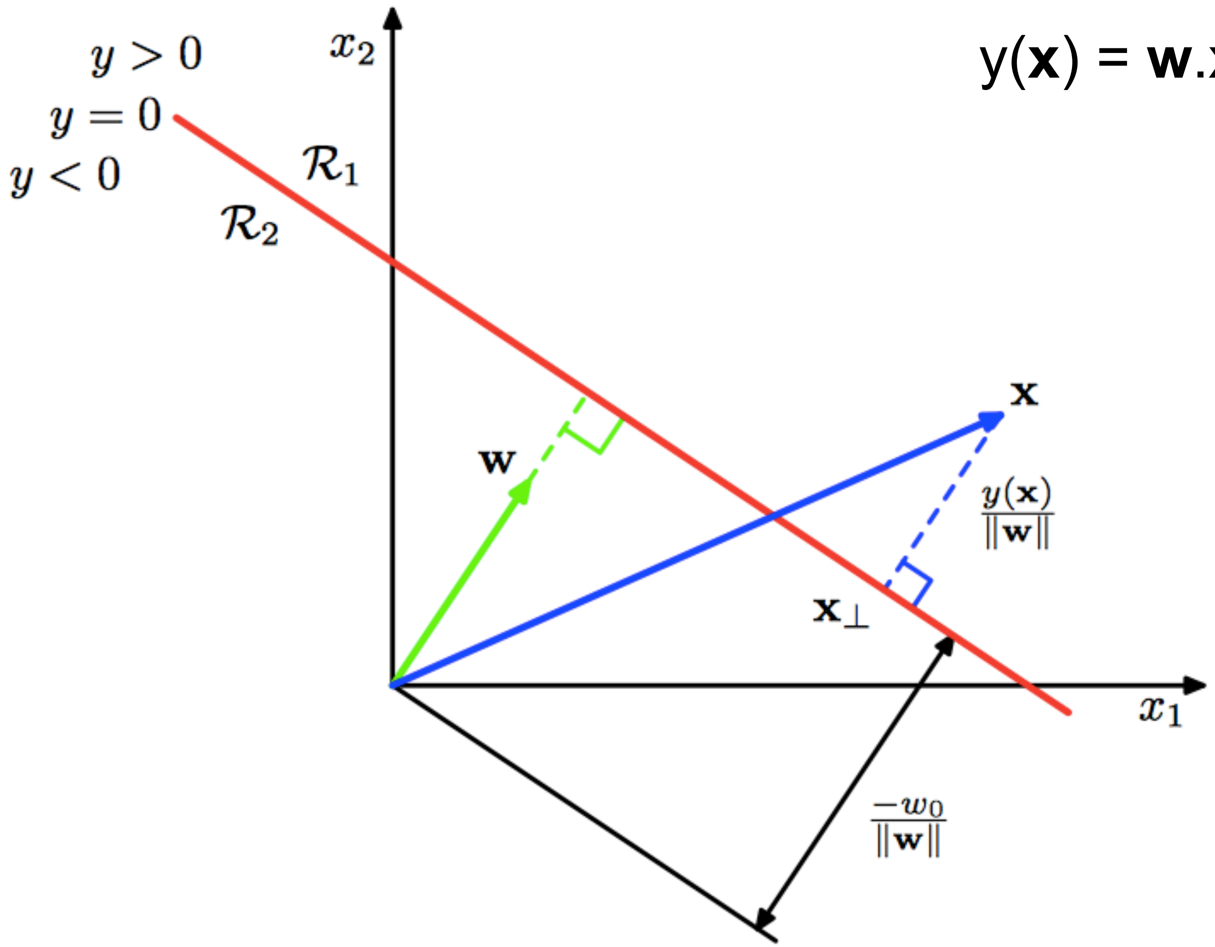
TRU CAO

HO CHI MINH CITY UNIVERSITY OF TECHNOLOGY
AND JOHN VON NEUMANN INSTITUTE

OUTLINE

- Review of analytical geometry
- Maximum margin classifiers
- Optimization using Langrage multipliers
- Kernel trick for non-linearly separable data
- Soft-margin SVMs

REVIEW OF ANALYTICAL GEOMETRY

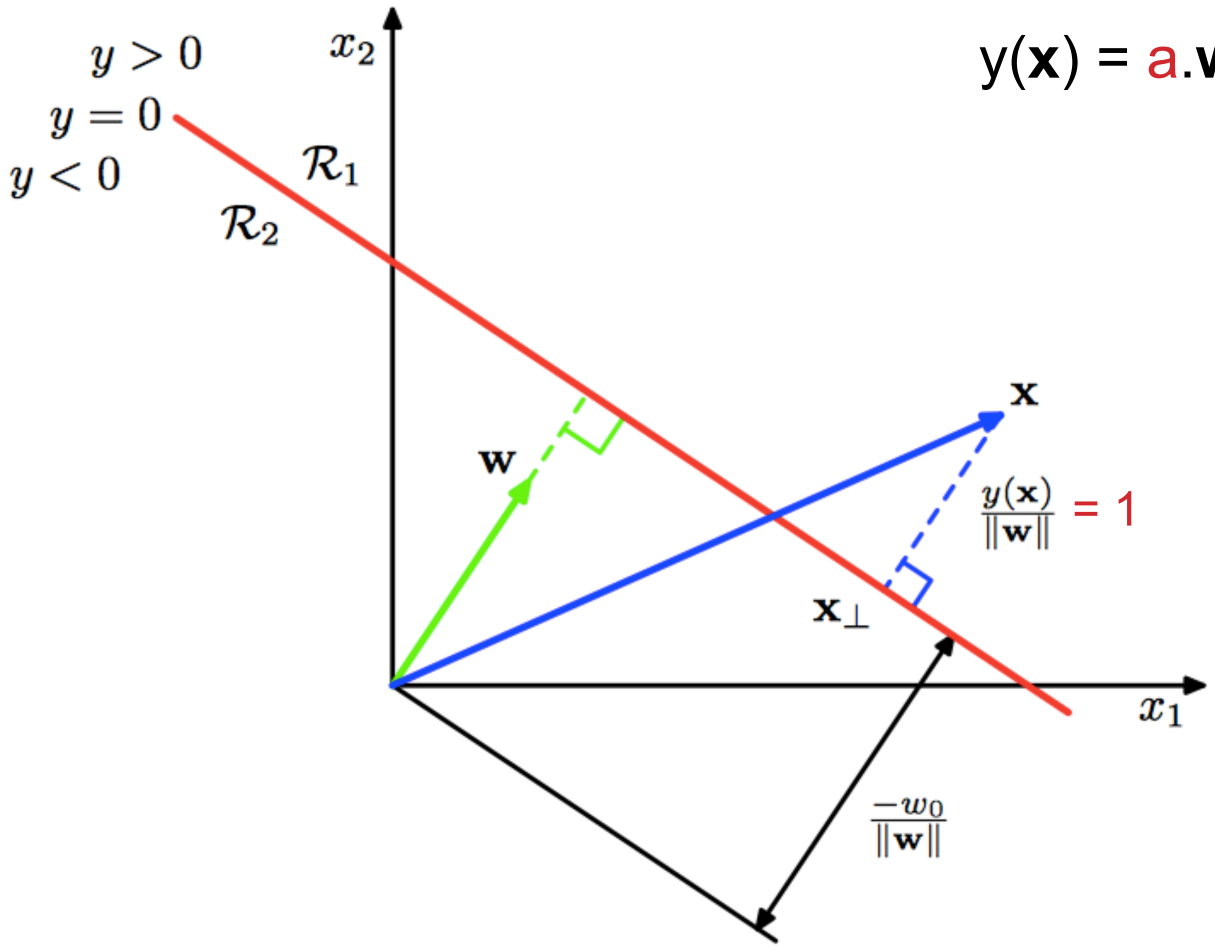


$y(\mathbf{x}) = \mathbf{w}.\mathbf{x} + w_0$

Bishop, C.M. (2006) Pattern Recognition and Machine Learning.

REVIEW OF ANALYTICAL GEOMETRY

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



Bishop, C.M. (2006) Pattern Recognition and Machine Learning.

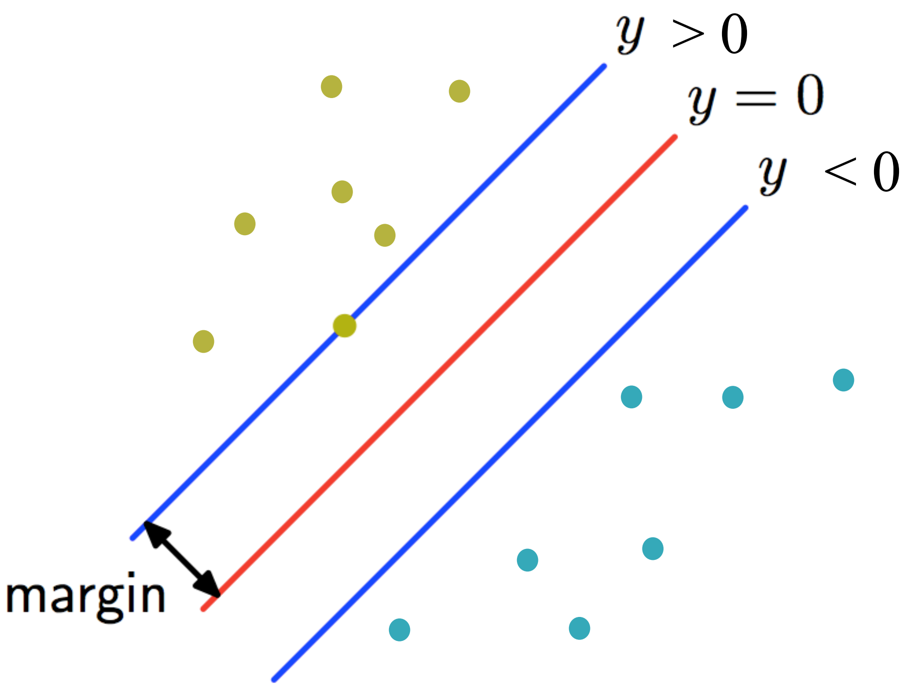
MAXIMUM MARGIN CLASSIFIERS

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

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

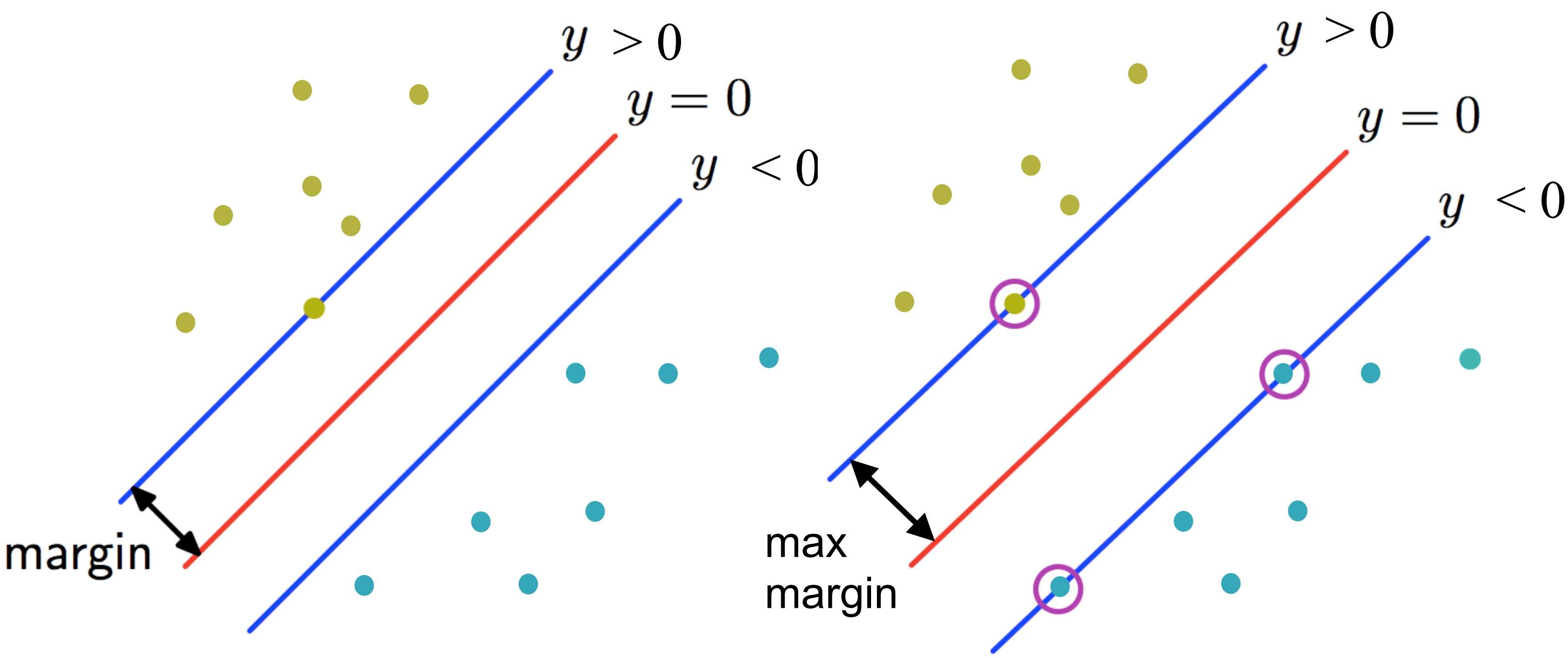
MAXIMUM MARGIN CLASSIFIERS

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



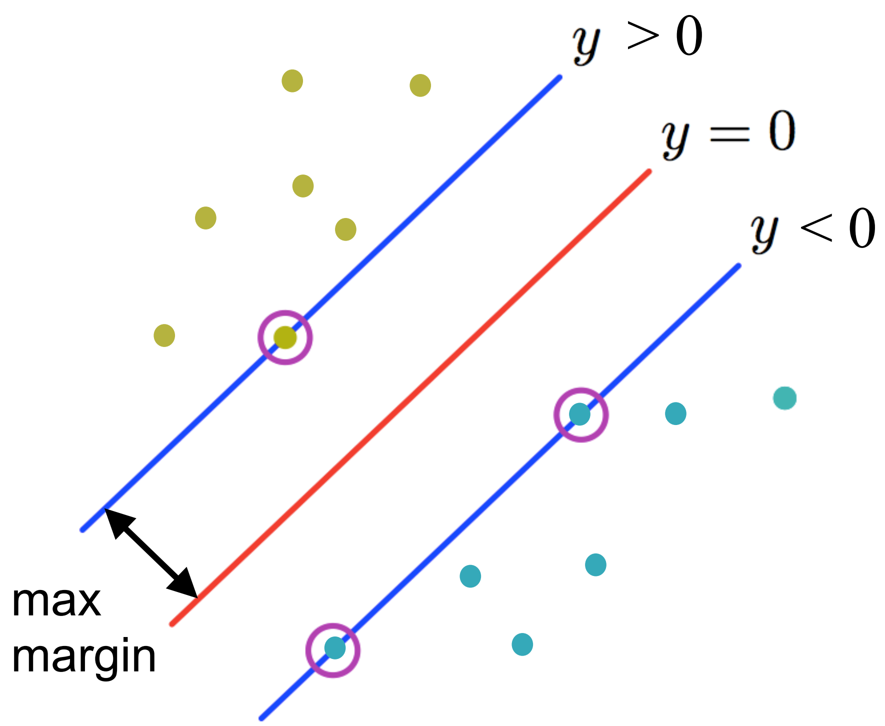
MAXIMUM MARGIN CLASSIFIERS

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



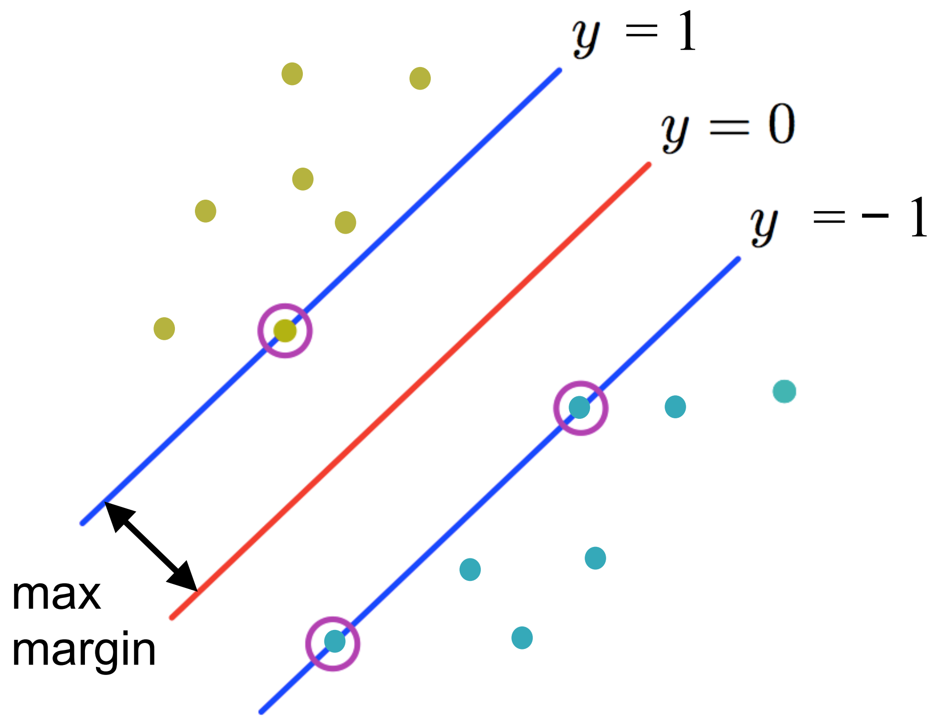
MAXIMUM MARGIN CLASSIFIERS

- Support vectors: samples at the two margins.



MAXIMUM MARGIN CLASSIFIERS

- Scaling the maximum margin to be 1:



MAXIMUM MARGIN CLASSIFIERS

- Signed distance between the decision boundary and a sample $\mathbf{x}_n$:

$$\frac{y(\mathbf{x}_n)}{||\mathbf{w}||}$$

MAXIMUM MARGIN CLASSIFIERS

- Signed distance between the decision boundary and a sample $\mathbf{x}_n$:

$$\frac{y(\mathbf{x}_n)}{||\mathbf{w}||}$$

- Absolute distance between the decision boundary and a sample $\mathbf{x}_n$:

$$\frac{t_n.y(\mathbf{x}_n)}{||\mathbf{w}||}$$

$t_n = +1$ iff $y(\mathbf{x}_n) > 0$ and $t_n = -1$ iff $y(\mathbf{x}_n) < 0$

MAXIMUM MARGIN CLASSIFIERS

- Maximum margin:

$$\operatorname{argmax}_{\mathbf{w},b}\left\{\frac{1}{||\mathbf{w}||}\min_n(t_n.(\mathbf{w}.\mathbf{x}_n + b))\right\}$$

with the constraint:

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

MAXIMUM MARGIN CLASSIFIERS

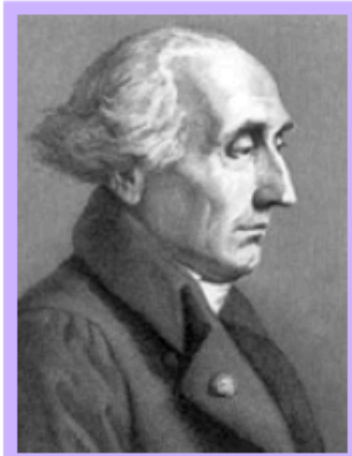
- To be optimized:

$$\text{argmin}_{\mathbf{w},b} \frac{1}{2} ||\mathbf{w}||^2$$

with the constraint:

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

OPTIMIZATION USING LAGRANGE MULTIPLIERS



Joseph-Louis Lagrange
1736–1813

Although widely considered to be a French mathematician, Lagrange was born in Turin in Italy. By the age of nineteen, he had already made important contributions mathematics and had been appointed as Professor at the Royal Artillery School in Turin. For many years, Euler worked hard to persuade Lagrange to move to Berlin, which he eventually did in 1766 where he succeeded Euler as Director of Mathematics at the Berlin Academy. Later he moved to Paris, narrowly escaping with his life during the French revolution thanks to the personal intervention of Lavoisier (the French chemist who discovered oxygen) who himself was later executed at the guillotine. Lagrange made key contributions to the calculus of variations and the foundations of dynamics.

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Problem:

$$\text{argmax}_{\mathbf{x}}\ f(\mathbf{x})$$

with the constraint:

$$g(\mathbf{x}) = 0$$

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Solution is the stationary point of the Lagrange function:

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

such that:

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

and

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

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Example:

$$f(\mathbf{x}) = 1 - u^2 - v^2$$

with the constraint:

$$g(\mathbf{x}) = u + v - 1 = 0$$

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Lagrange function:

$$L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda.g(\mathbf{x}) = (1 - u^2 - v^2) + \lambda.(u + v - 1)$$

$$\partial L(\mathbf{x}, \lambda)/\partial u = \partial f(\mathbf{x})/\partial u + \lambda.\partial g(\mathbf{x})/\partial u = -\ 2u + \lambda = 0$$

$$\partial L(\mathbf{x}, \lambda)/\partial v = \partial f(\mathbf{x})/\partial v + \lambda.\partial g(\mathbf{x})/\partial v = -\ 2v + \lambda = 0$$

$$\partial L(\mathbf{x}, \lambda)/\partial \lambda = g(\mathbf{x}) = u + v - 1 = 0$$

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

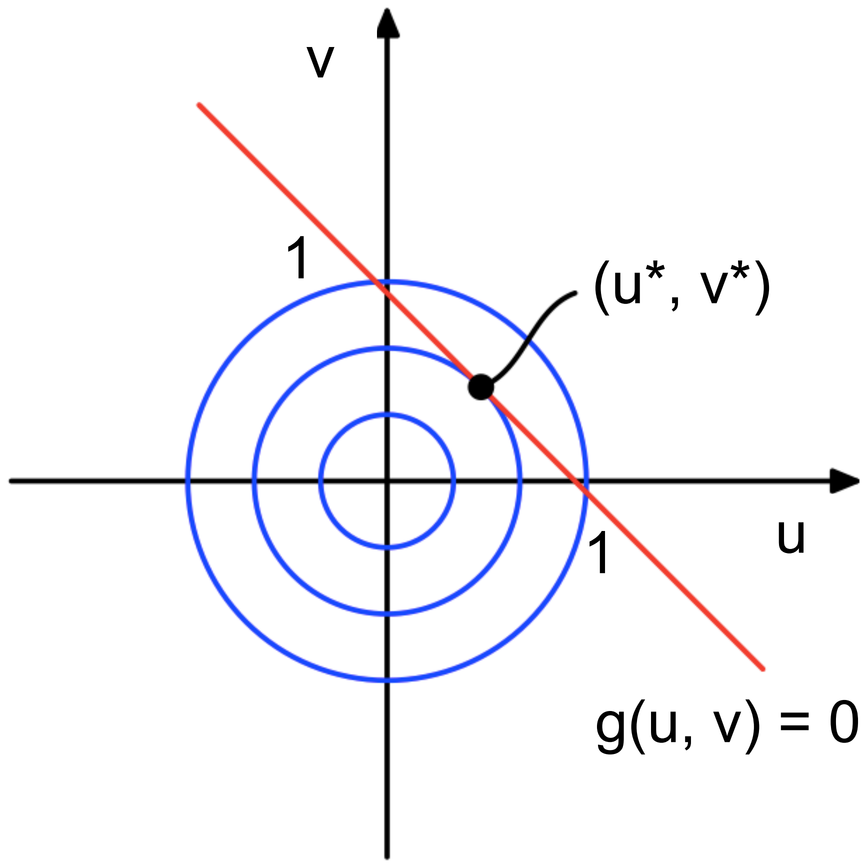
OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Example:

  $f(\mathbf{x}) = 1 - u^2 - v^2$

  with the constraint:

  $g(\mathbf{x}) = u + v - 1 = 0$



OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Problem:

$$\text{argmax}_{\mathbf{x}}\ f(\mathbf{x})$$

with the inequality constraint:

$$g(\mathbf{x}) \geq 0$$

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Solution is the stationary point of the Lagrange function:

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

such that:

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

and

$$g(\mathbf{x}) \geq 0$$

$$\lambda \geq 0$$

$$\lambda.g(\mathbf{x}) = 0$$

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- To be optimized:

$$\text{argmin}_{\mathbf{w},b} \frac{1}{2} ||\mathbf{w}||^2$$

with the constraint:

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

- Lagrange function for maximum margin classifier:

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

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

$$a_n \geq 0$$

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

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Lagrange function for maximum margin classifier:

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

- Solution for **w**: $\partial L(\mathbf{w}, b, \mathbf{a})/\partial \mathbf{w} = 0$

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

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

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Lagrange function for maximum margin classifier:

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

- Solution for **a**: dual representation to be minimized

$$L^*(\mathbf{a}) = \sum_{n=1..N} a_n - \frac{1}{2} \sum_{n=1..N} \sum_{n=1..M} a_n.a_m.t_n.t_m.\mathbf{x}_n.\mathbf{x}_m$$

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:

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

- Solution for **a**: dual representation to be minimized

$$L^*(\mathbf{a}) = \sum_{n=1..N} a_n - \frac{1}{2} \sum_{n=1..N} \sum_{n=1..M} a_n.a_m.t_n.t_m.\mathbf{x}_n.\mathbf{x}_m$$

Why optimization via dual representation?

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Lagrange function for maximum margin classifier:

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

- Solution for **a**: dual representation to be minimized

$$L^*(\mathbf{a}) = \sum_{n=1..N} a_n - \frac{1}{2} \sum_{n=1..N} \sum_{n=1..M} a_n.a_m.t_n.t_m.\mathbf{x}_n.\mathbf{x}_m$$

Why optimization via dual representation?

- Sparsity: $a_n = 0$ if $\mathbf{x}_n$ is not a support vector.

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Lagrange function for maximum margin classifier:

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

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

- Solution for b:

$$b = \frac{1}{|S|} \sum_{n\in S} (t_n - \sum_{m\in S} a_m.t_m.\mathbf{x}_m.\mathbf{x}_n)$$

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

OPTIMIZATION USING LAGRANGE MULTIPLIERS

- Classification:

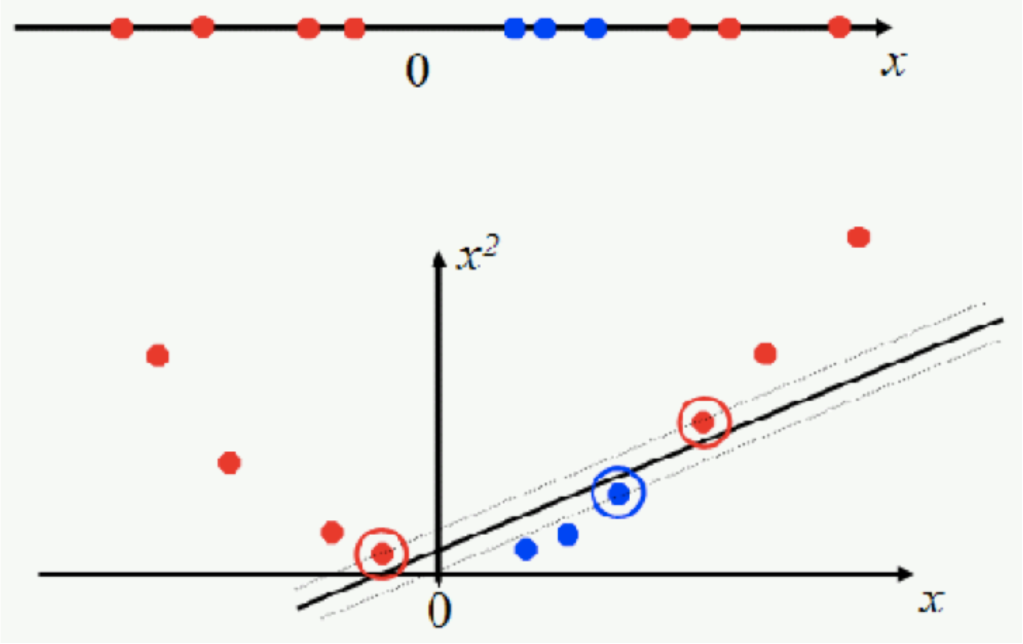
$$y(\mathbf{x}) = \mathbf{w}.\mathbf{x} + b = \Sigma_{n=1..N}\, a_n.t_n.\mathbf{x}_n.\mathbf{x} + b$$

$$y(\mathbf{x}) > 0 \Rightarrow +1$$

$$y(\mathbf{x}) < \Rightarrow -1$$

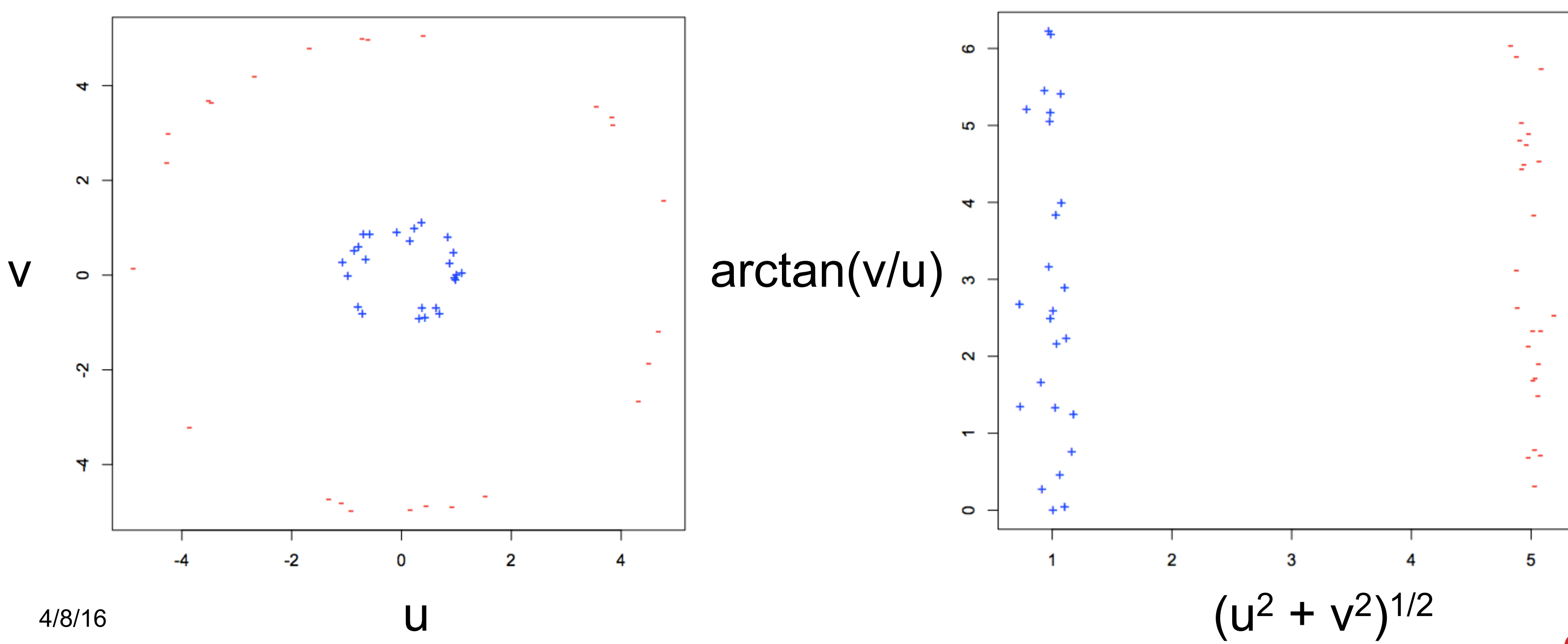
KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- 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

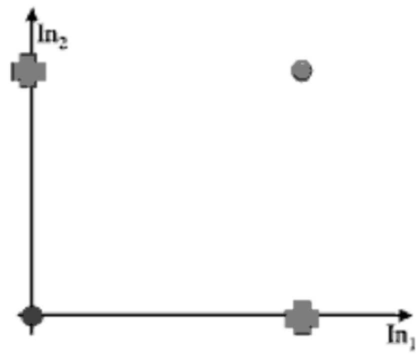
- Example 2:
  - Original space: (u, v)
  - New space: $((u^2 + v^2)^{1/2}, \arctan(v/u))$



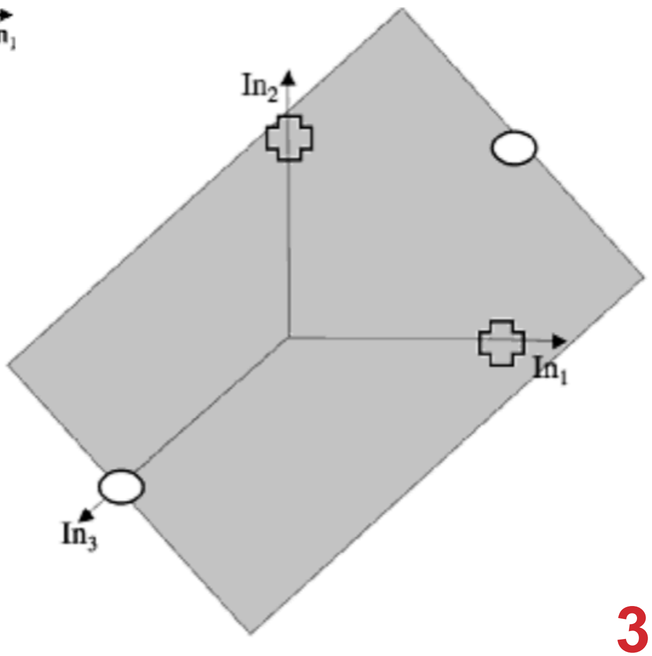
KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- Example 3: XOR function

In_1	In_2	t
0	0	0
0	1	1
1	0	1
1	1	0



In_1	In_2	In_3	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:

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

KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- Classification in the new space:

$$y(\mathbf{x}) = \mathbf{w}.\Phi(\mathbf{x}) + b = \Sigma_{n=1..N}\, a_n.t_n.\Phi(\mathbf{x}_n).\Phi(\mathbf{x}) + b$$

- Computational complexity of $\Phi(\mathbf{x}_n).\Phi(\mathbf{x})$ is high due to the high dimension of $\Phi(.)$.

KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- Classification in the new space:

$$y(\mathbf{x}) = \mathbf{w}.\Phi(\mathbf{x}) + b = \Sigma_{n=1..N}\, a_n.t_n.\Phi(\mathbf{x}_n).\Phi(\mathbf{x}) + b$$

- Computational complexity of $\Phi(\mathbf{x}_n)^T.\Phi(\mathbf{x})$ is high due to the high dimension of $\Phi(.)$.

- Kernel trick:

$$\Phi(\mathbf{x}_n).\Phi(\mathbf{x}_m) = K(\mathbf{x}_n, \mathbf{x}_m)$$

KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- A typical kernel function:

$$K(\mathbf{u}, \mathbf{v}) = (1 + \mathbf{u}.\mathbf{v})^2$$

$$\Phi((u_1, u_2, \ldots, u_d)) = (1, \sqrt{2}u_1, \sqrt{2}u_2, \ldots, \sqrt{2}u_d,$$
$$\sqrt{2}\,u_1.u_2, \sqrt{2}u_1.u_3, \ldots, \sqrt{2}\,u_{d-1}.u_d,$$
$$u_1^2, u_2^2, \ldots, u_d^2)$$

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

$$\Phi(\mathbf{u}).\Phi(\mathbf{v}) = K(\mathbf{u}, \mathbf{v})$$

KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- Is $\Phi(\mathbf{x})$ guaranteed to be linearly separable?

KERNEL TRICK FOR NON-LINEARLY SEPARABLE DATA

- Soft-margin SVM is introduced.

HOMEWORK

- Reading Appendix E about Lagrange multipliers in Bishop (2006), Pattern Recognition and Machine Learning.
- Verify all computations and equations in the lecture slides.