



MÁY HỌC NÂNG CAO (Máy học véc-to hỗ trợ - SVM)

Phạm Nguyên Khang
pnkhang@cit.ctu.edu.vn



CANTHO UNIVERSITY

Nội dung trình bày

- Giới thiệu
- SVM cho bài toán phân lớp
- Giải thuật SVM
- SVM cho bài toán đa lớp
- SVM cho bài toán hồi quy (xem giáo trình)
- SVM cho bài toán phát hiện phần tử cá biệt (xem giáo trình)
- Proximal SVM
- Bài tập



CANTHO UNIVERSITY

Giới thiệu

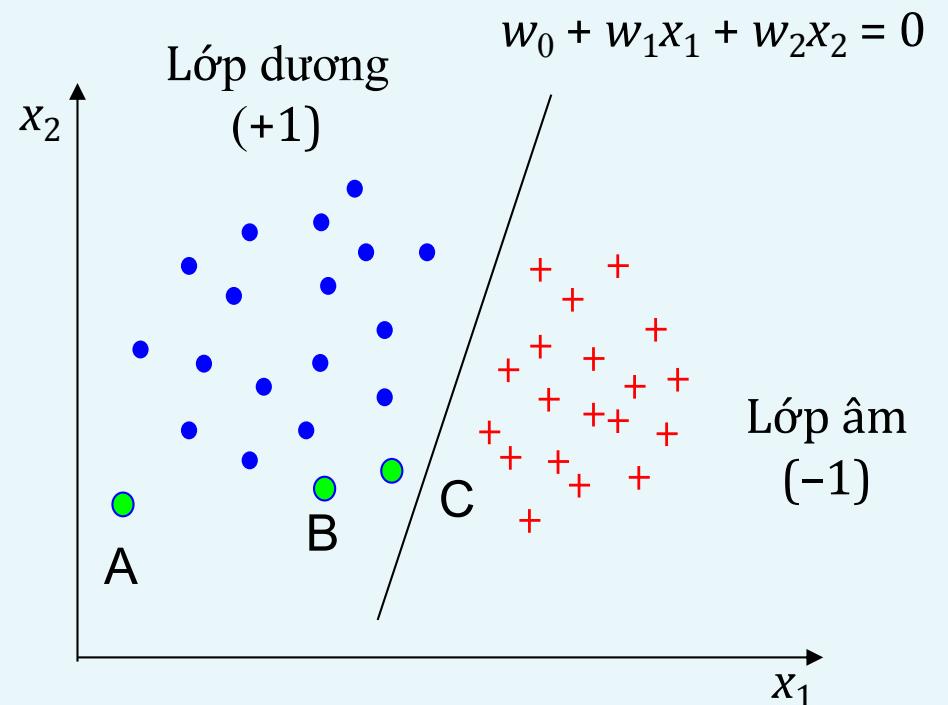
- Máy học véc-tơ hỗ trợ (support vector machines – SVM) được Vapnik nghiên cứu từ những năm 1965
- Giải thuật được phát triển mạnh vào những năm 1990
- Là công cụ hữu hiệu và phổ biến của lĩnh vực máy học, nhận dạng và khai mỏ dữ liệu.
- Áp dụng thành công trong: nhận dạng mặt người, phân loại văn bản, phân loại bệnh ung thư, ...
- Có thể áp dụng phương pháp hàm nhân (kernel method).



CANTHO UNIVERSITY

SVM cho bài toán phân lớp

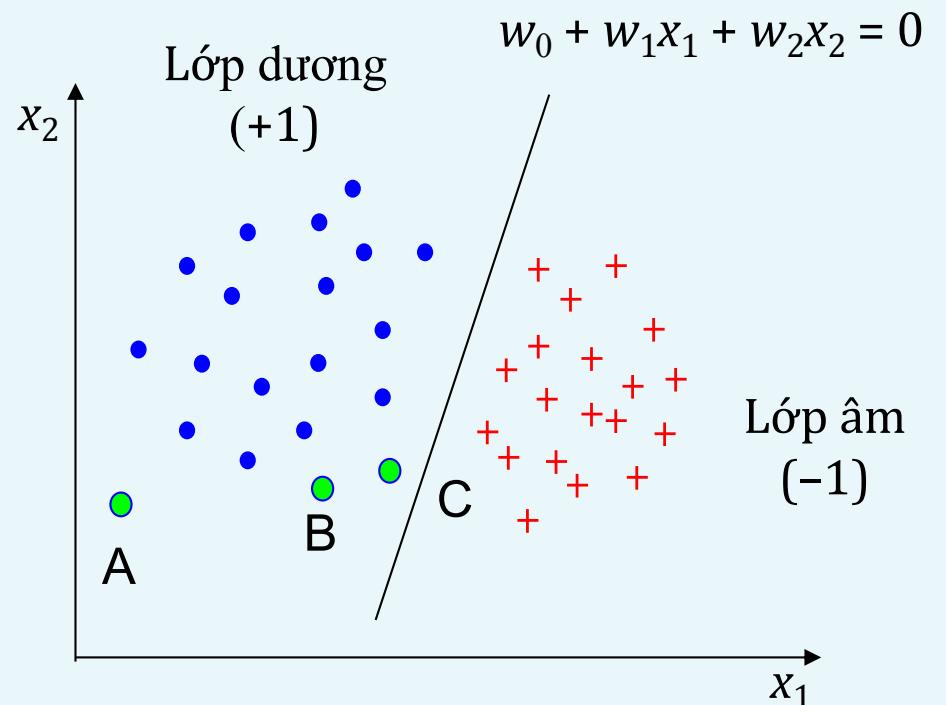
- Bài toán phân lớp nhị phân (hai chiều)
 - Lớp dương (+1)
 - Lớp âm (-1)
 - Phân lớp 3 phần tử mới: A, B, C ?
- Độ tin cậy phân lớp ?





SVM cho bài toán phân lớp

- Bài toán phân lớp nhị phân
 - Lớp dương (+1)
 - Lớp âm (-1)
 - Phân lớp 3 phần tử mới: A, B, C ?
- Độ tin cậy phân lớp ?
 - Càng cách xa đường biên càng an toàn
 - Tìm đường biên như thế nào để an toàn ?

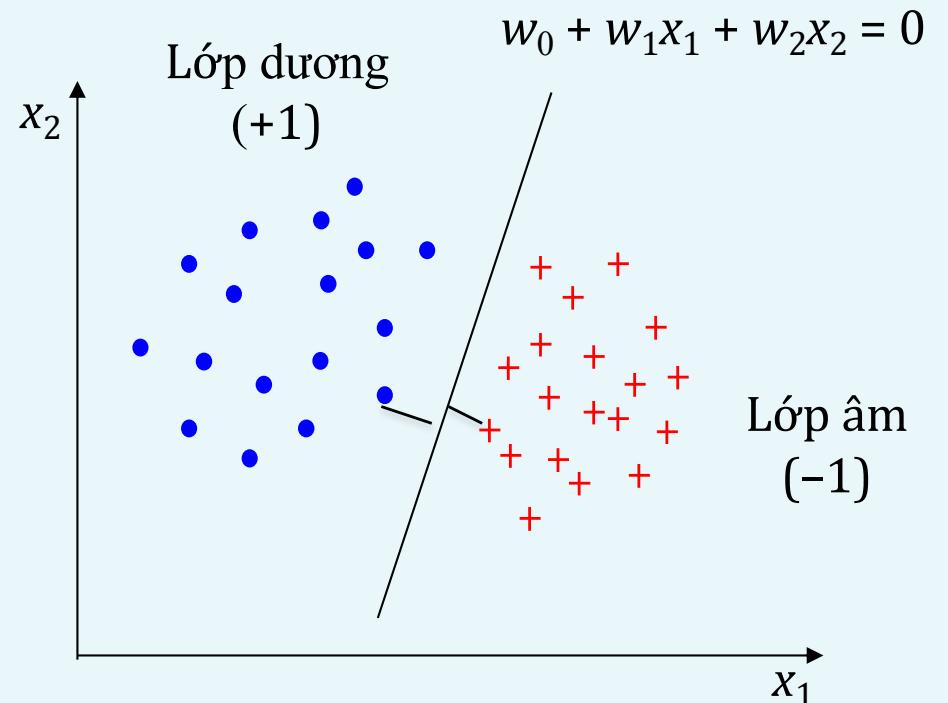




CANTHO UNIVERSITY

SVM cho bài toán phân lớp

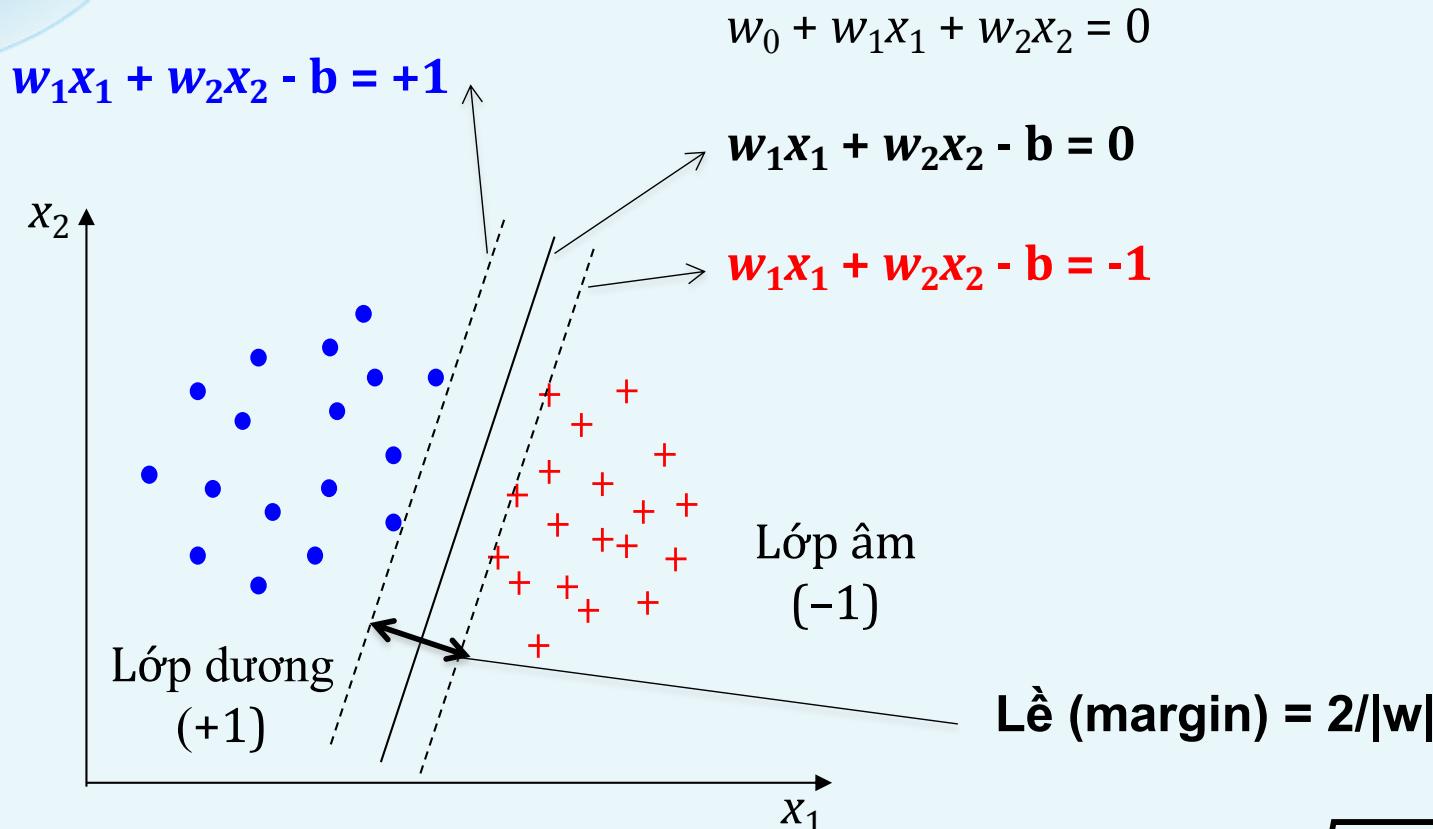
- Tìm đường biên như thế nào để an toàn ?
 - Trực quan: tìm đường biên chia dữ liệu thành hai phía, mỗi lớp ở 1 phía (**phân lớp đúng**) và các phần tử càng cách xa đường biên càng tốt (**tin cậy cao**).





CANTHO UNIVERSITY

SVM cho bài toán phân lớp



Các phần tử càng cách xa đường biên
càng tốt ~ Lề càng lớn càng tốt.

$$|w| = \sqrt{w_1^2 + w_2^2}$$



CANTHO UNIVERSITY

SVM cho bài toán phân lớp

- Lớp dương: $y_i = 1$
 - $w_1x_1 + w_2x_2 - b \geq +1$
- Lớp âm: $y_i = -1$
 - $w_1x_1 + w_2x_2 - b \leq -1$
- Nhân 2 vế cho y_i :
 - $y_i(w_1x_1 + w_2x_2 - b) \geq +1$
- Bài toán trở thành tìm $\mathbf{w} = (w_1, w_2)$ và b sao cho:
 - $2/\|\mathbf{w}\| \rightarrow \max$ (hay $|\mathbf{w}|^2 \rightarrow \min$)
 - với ràng buộc: $y_i(w_1x_1 + w_2x_2 - b) \geq +1$



CANTHO UNIVERSITY

SVM cho bài toán phân lớp

- Tổng quát cho trường hợp n chiều
 - Tìm $\mathbf{w} = (w_1, w_2, \dots, w_n)$ và b sao cho:
 - $\frac{1}{2}|\mathbf{w}|^2 \rightarrow \min$
 - với ràng buộc: $y_i(\mathbf{w}^T \mathbf{x}_i) \geq +1$

$$|\mathbf{w}| = \sqrt{w_1^2 + w_2^2 + \cdots + w_n^2}$$

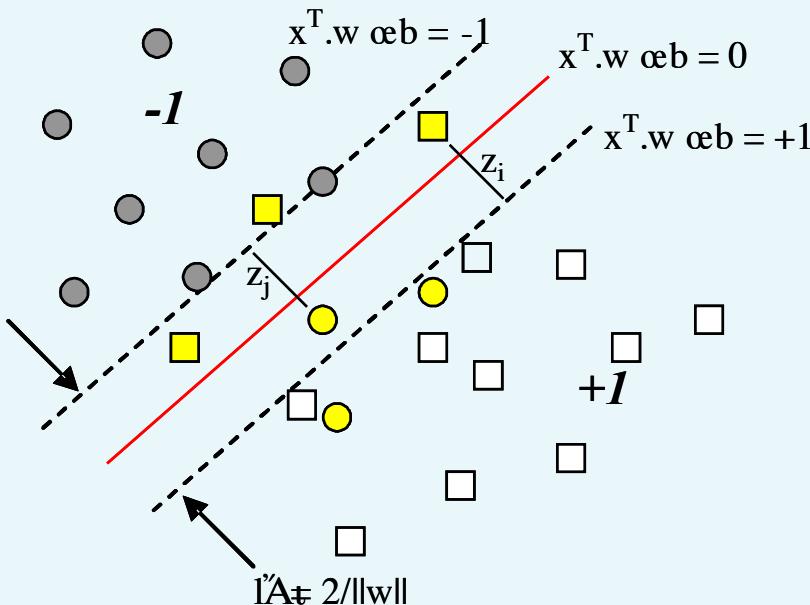
- Đây chính là bài toán **quy hoạch toàn phương** (Quadratic programming)
- Giải bài toán này ta thu được $\mathbf{w} = (w_1, w_2, \dots, w_n)$ và b
- Dự báo nhãn cho phần tử mới:
 - Nhãn = $\text{sign}(\mathbf{w}^T \mathbf{x} - b)$



CANTHO UNIVERSITY

SVM cho bài toán phân lớp

- Trường hợp dữ liệu không khả tách tuyến tính:
 - Chấp nhận lỗi huấn luyện
 - Bài toán: tìm w, b sao cho lề lớn nhất và lỗi huấn luyện nhỏ nhất





CANTHO UNIVERSITY

SVM cho bài toán phân lớp

- Giải bài toán quy hoạch toàn phương
 - Phương pháp nhân tử Larange

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j x_i \cdot x_j - \sum_{i=1}^m \alpha_i \rightarrow \min$$

- Với ràng buộc

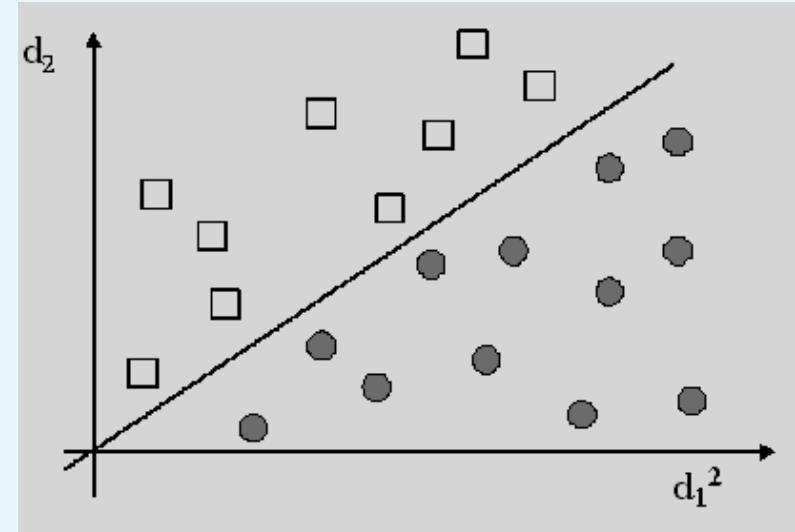
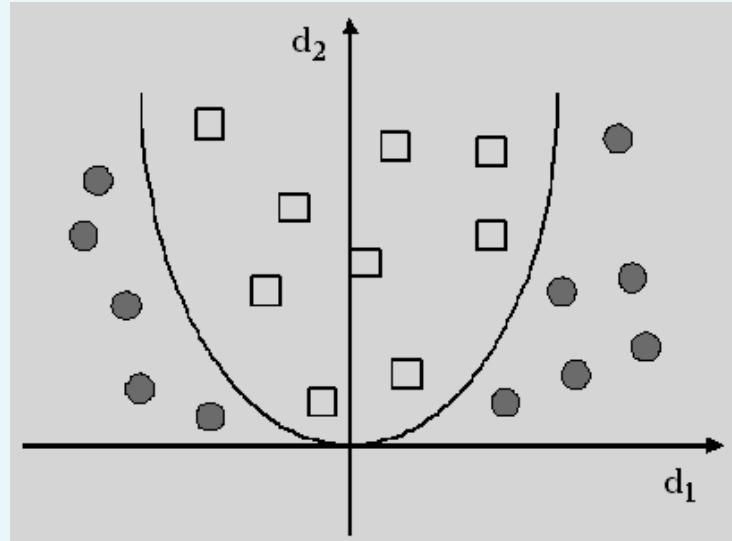
$$\sum_{i=1}^m \alpha_i y_i = 0$$

$$\alpha_i \geq 0$$



CANTHO UNIVERSITY

SVM và phương pháp hàm nhân



Áp dụng SVM tuyến tính
trên không gian mới

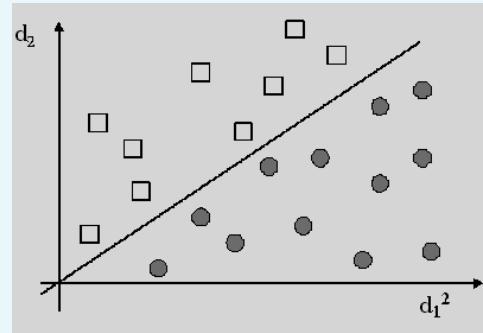
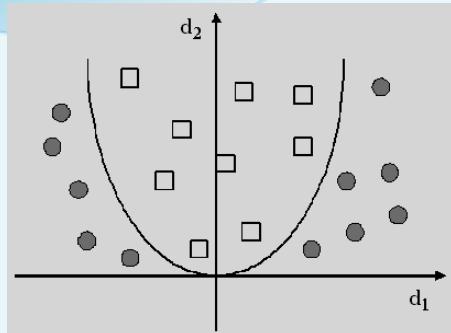
Ánh xạ phi tuyến Φ
 $X \rightarrow \Phi(X)$

Hầu như không thể tìm được $\Phi(X) !!!$



CANTHO UNIVERSITY

SVM và phương pháp hàm nhân



Ánh xạ phi tuyến Φ
 $X \rightarrow \Phi(X)$

Tính toán trên
X

Tính toán trên
 $(X_i \cdot X_j)$

Tính toán trên
 $\Phi(X)$

Tính toán trên
 $(\Phi(X_i) \cdot \Phi(X_j)) = K(X_i, X_j)$

Kernel trick:

Nếu mô hình không tính toán trực tiếp lên x mà tính toán thông qua các tích vô hướng dạng $(x_i x_j)$, ta có thể không cần ánh xạ Φ mà vẫn xử lý được dữ liệu phi tuyến.

Kernel function:

$K(X_i, X_j)$: hàm nhân được định nghĩa trong không gian gốc = tích vô hướng trong không gian trung gian (feature space)



CANTHO UNIVERSITY

SVM và phương pháp hàm nhân

- SVM tuyến tính trong không gian gốc

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j x_i \cdot x_j - \sum_{i=1}^m \alpha_i \rightarrow \min$$

- SVM tuyến tính trong không gian trung gian

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \Phi(x_i) \cdot \Phi(x_j) - \sum_{i=1}^m \alpha_i =$$

$$\frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^m \alpha_i \rightarrow \min$$



CANTHO UNIVERSITY

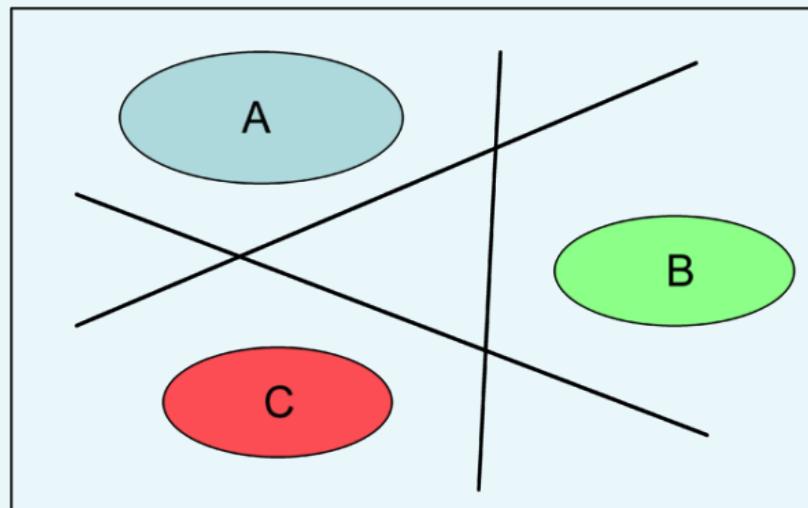
SVM và phương pháp hàm nhân

- Một số hàm nhân thông dụng:
 - Tuyến tính: $K(u, v) = u \cdot v$
 - Đa thức bậc d : $K(u, v) = (u \cdot v + c)^d$
 - Radial Basis Function (RBF):
 - $K(u, v) = \exp(-\gamma \|u - v\|^2)$
- Cài đặt trong libSVM:
 - `-s svm_type` (default 0): 0 (phân lớp), 2 (1 lớp), 3 (hồi quy)
 - `-t kernel_type` (default 2): 0 (tuyến tính), 1 (đa thức), 2 (RBF)
 - `-d degree` (default 3): bậc của hàm nhân đa thức
 - `-g gamma` (default 1/#attr): tham số γ của hàm nhân RBF
 - `-c cost` (default 1) : hằng số c

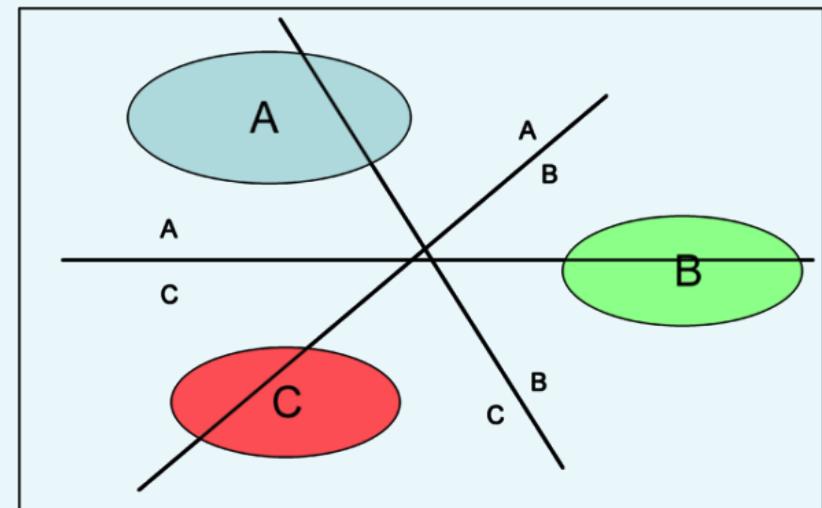


CANTHO UNIVERSITY

SVM cho bài toán đa lớp



1 chống lại tất cả
(1 vs all)



1 - 1
(1 vs 1)



CANTHO UNIVERSITY

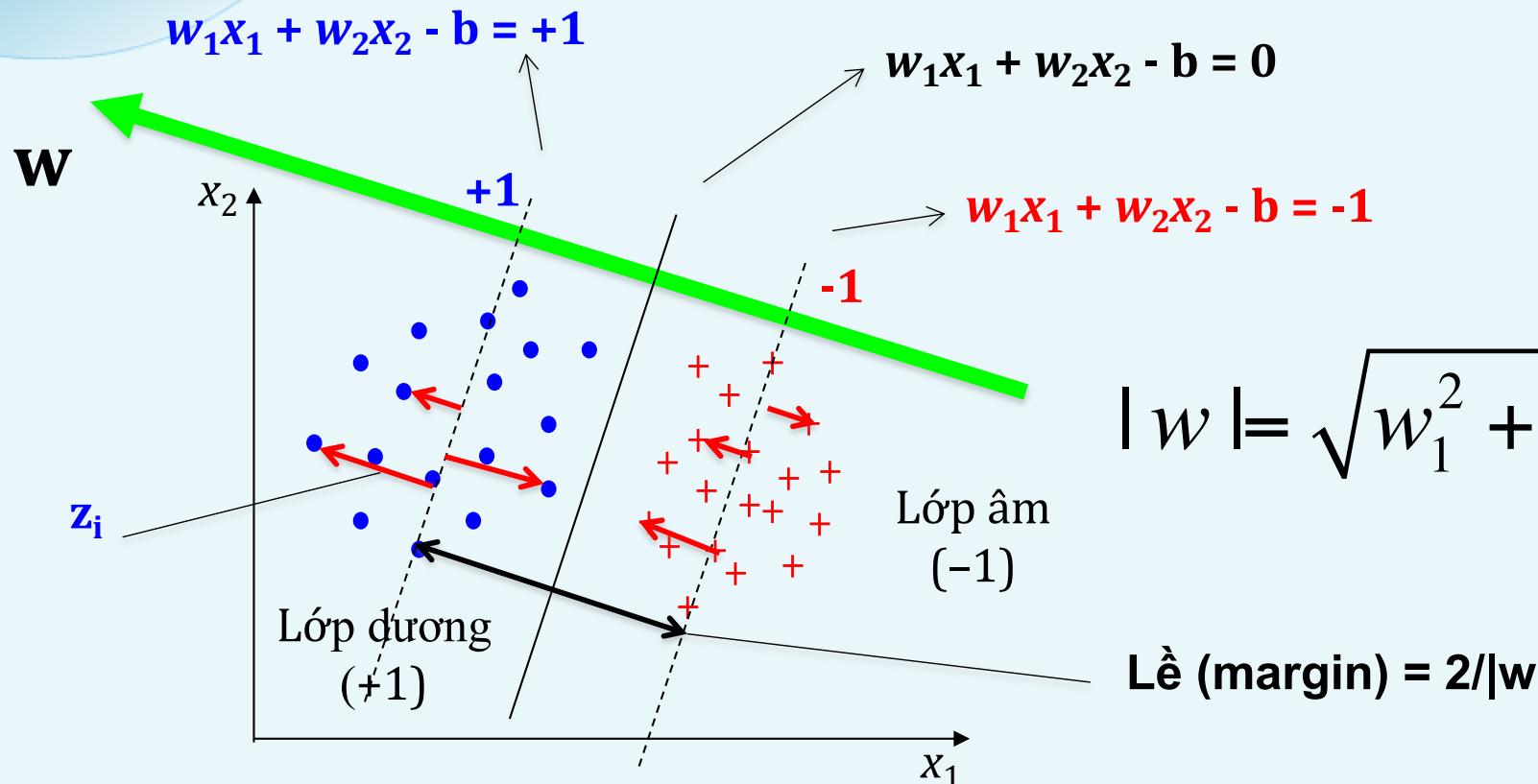
Proximal SVM

- Do Fung và Mangasarian đề xuất
- Thay đổi hai siêu phẳng hỗ trợ
- Thay đổi độ lớn của lề
- Bài toán quy hoạch toàn phương \rightarrow bài toán giải hệ phương trình tuyến tính



CANTHO UNIVERSITY

Proximal SVM



$$|w| = \sqrt{w_1^2 + w_2^2 + b^2}$$

$$\text{Lè (margin)} = 2/|w|$$

Lè càng lớn càng tốt và tổng khoảng cách từ các phân tử đến siêu phẳng hỗ trợ càng nhỏ càng tốt



CANTHO UNIVERSITY

Proximal SVM

$$c \frac{1}{2} \sum_i z_i^2 + \frac{1}{2} (w_1^2 + w_2^2 + b^2) \rightarrow \min$$

với ràng buộc:

$$y_i(w^T x_i - b) + z_i = 1$$



CANTHO UNIVERSITY

Proximal SVM

- A: ma trận dữ liệu mxn, mỗi hàng là 1 phần tử
- D: ma trận đường chéo chính chứa nhãn của các phần tử mxm
- e: véc-tơ cột chỉ chứa toàn số 1: mx1
- Giải thuật Proximal SVM:
 - Xây dựng ma trận $H = D[A - e]$
 - Tính $u = c(I - H(1/c + H'H)^{-1}H'e)$
 - Tính $w = A'Du$
 - Tính $b = -e'Du$
- Phân lớp phần tử mới:
 - Nhãn = $\text{sign}(w^T x - b)$



CANTHO UNIVERSITY

Proximal SVM

- Tham khảo thêm:
 - <http://research.cs.wisc.edu/dmi/svm/psvm/>

