MÁY HỌC VECTOR HỖ TRỢ
SUPPORT VECTOR MACHINE

NỘI DUNG

1. Giới thiệu SVM
2. Tại sao chọn SVM
3. Đặt vấn đề
4. Bài toán phân 2 lớp với SVM
5. So cánh và cải tiền SVM

1. Giới thiệu SVM

Giới thiệu	• Phương pháp Support vector machines.
Lịch Sử	• Lý thuyết học thống kê do Vapnik và Chervonekis xây dựng năm 1995
Định nghĩa	• Dựa trên nền tảng lý thuyết thống kê, sử dụng cho phân lớp dữ liệu.
Ứng Dụng	• Nhận dạng, phân tích dữ liệu, ký tự,…

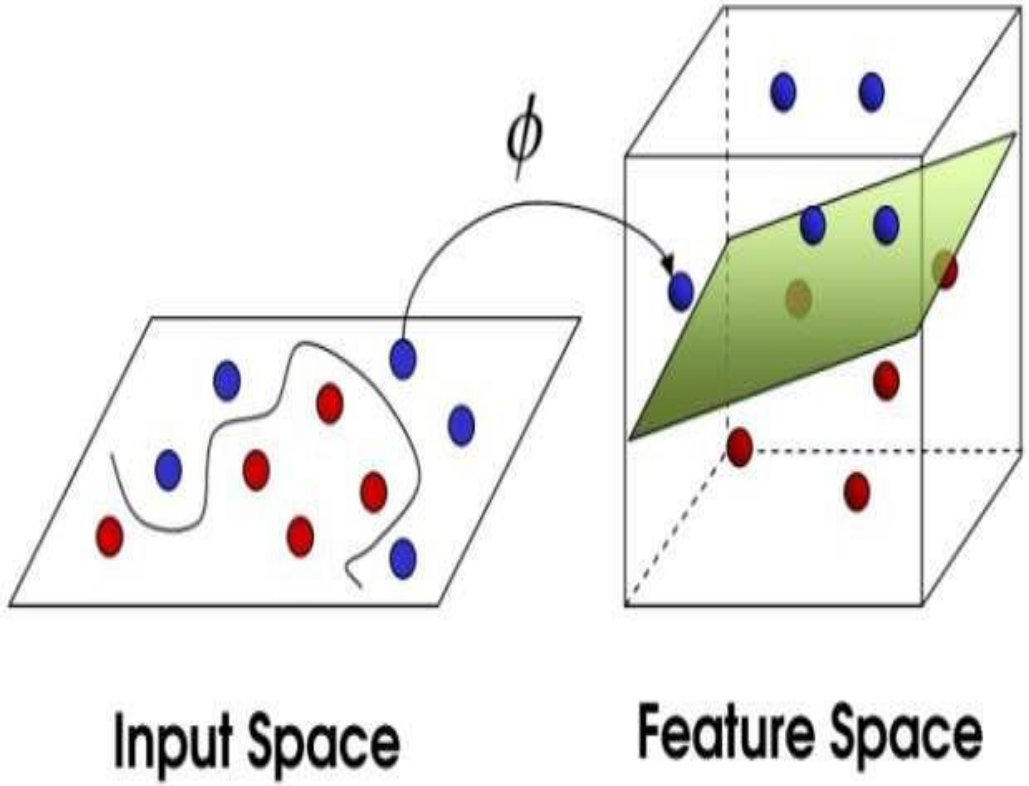
2. Tại sao chọn SVM ?

1. Hiệu quả giải quyết bài toán dữ liệu có số chiều lớn (ảnh của dữ liệu biểu diễn gene, protein, tế bào)
2. Giải quyết vấn đề overfitting rất tốt (dữ liệu có nhiễu và tách rời nhóm hoặc dữ liệu hoặc dữ liệu huấn luyện quá ít)
3. Là phương pháp phân lớp nhanh.
4. Có hiệu suất tổng hợp tốt và hiệu suất tính toán cao.

3. Đặt vấn đề

- Cho trước một số điểm dữ liệu cùng với nhãn của chúng thuộc một trong hai lớp cho trước.
- Mục tiêu của thuật toán là xác định xem một điểm dữ liệu *mới* sẽ được thuộc về lớp nào.
- Mỗi điểm dữ liệu được biểu diễn dưới dạng một vector p-chiều và ta muốn chia tách hai lớp dữ liệu bằng một siêu phẳng $p - 1$ chiều.

➢ Đây gọi là **phân loại tuyến tính**.



Cơ sở lý thuyết

Cho tập huấn luyện D gồm n điểm có dạng

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) \mid \mathbf{x}_i \in \mathbb{R}^p,\ y_i \in \{-1, 1\}\}_{i=1}^{n}$$

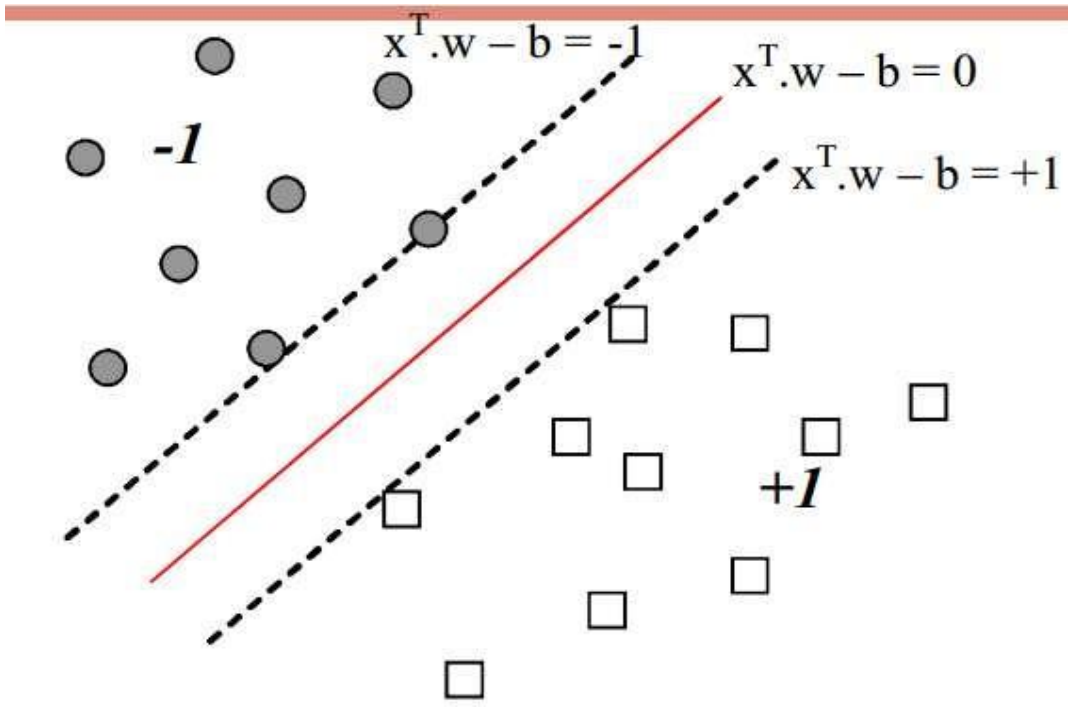
$\mathbf{y_i}$: mang giá trị 1 và -1, xác định lớp của điểm $\mathbf{x_i}$.

$\mathbf{x_i}$: Là một vector thực nhiều chiều (p chiều).

w: Là một vectơ pháp tuyến của siêu phẳng.

PT siêu phẳng chứa vector $\vec{x_i}$ trong không gian:

$$\vec{x_i}.\vec{w} + b = 0$$



Cơ sở lý thuyết

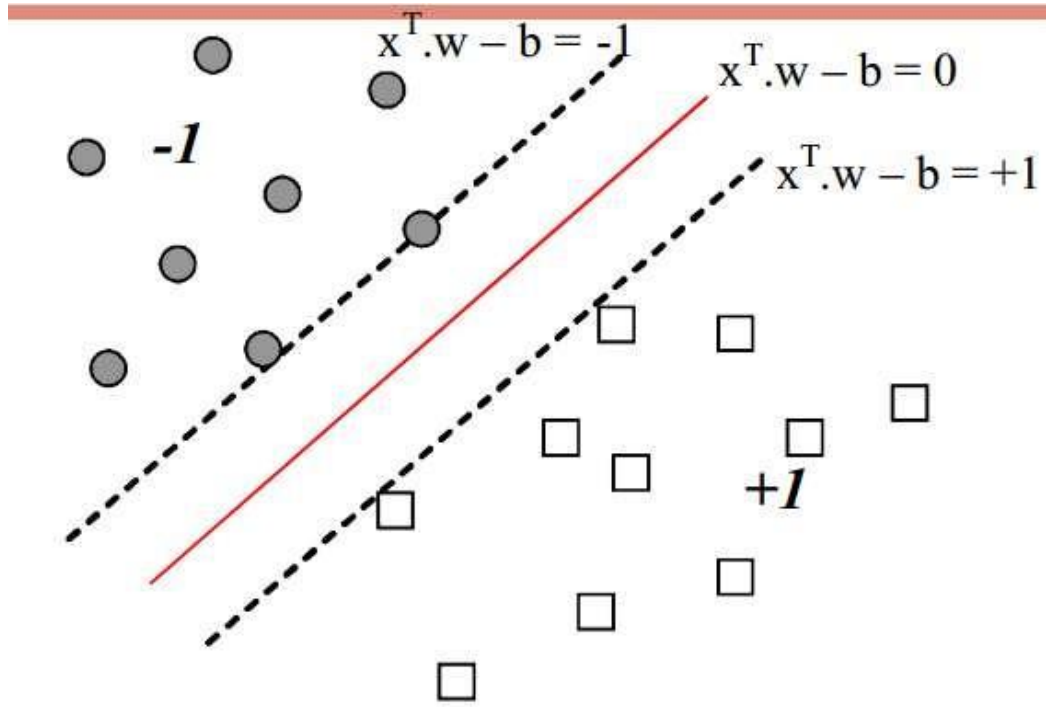
Đặt $f(\vec{x_i}) = \text{sign}(\vec{x_i}.\vec{w} + b) = \begin{cases} +1, \vec{x_i}\vec{w} + b > 0 \\ -1, \vec{x_i}.\vec{w} + b < 0 \end{cases}$

Như vậy, $f(\vec{x_i})$ biểu diễn sự phân lớp của $\vec{x_i}$ vào hai lớp như nêu trên.

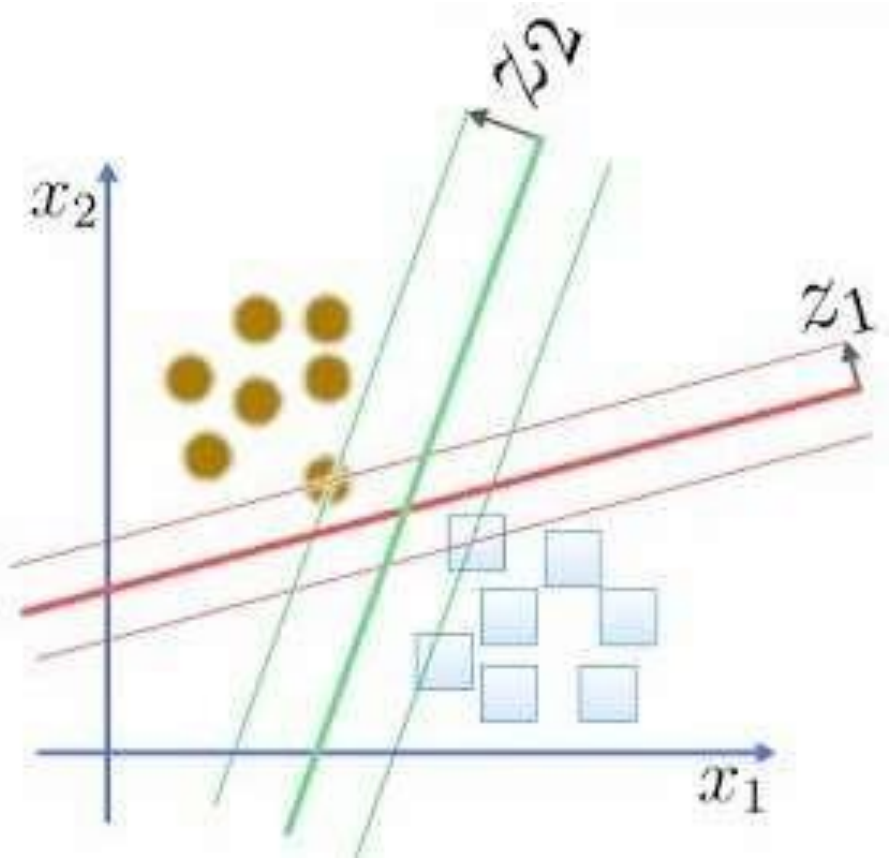
Ta nói :

$y_i = +1$ nếu $\vec{x_i}$ thuộc lớp I

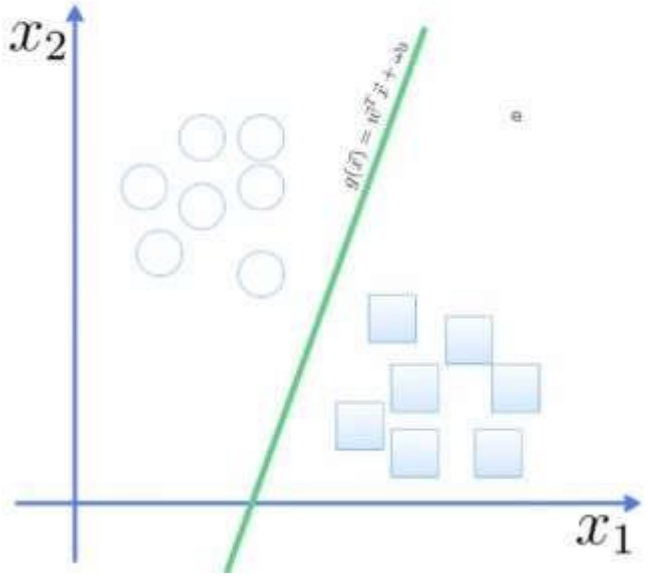
$y_i = -1$ nếu $\vec{x_i}$ thuộc lớp II.



Ví dụ



$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class } 1$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class } 2$$



4. Bài toán phân 2 lớp với SVM

TH1: Tập D phân chia tuyến tính không nhiễu

Đặt $f(\vec{x_i}) = \text{sign}(\vec{x_i}.\vec{w} + b) = \begin{cases} +1, \vec{x_i}.\vec{w} + b > 0 \\ -1, \vec{x_i}.\vec{w} + b < 0 \end{cases}$

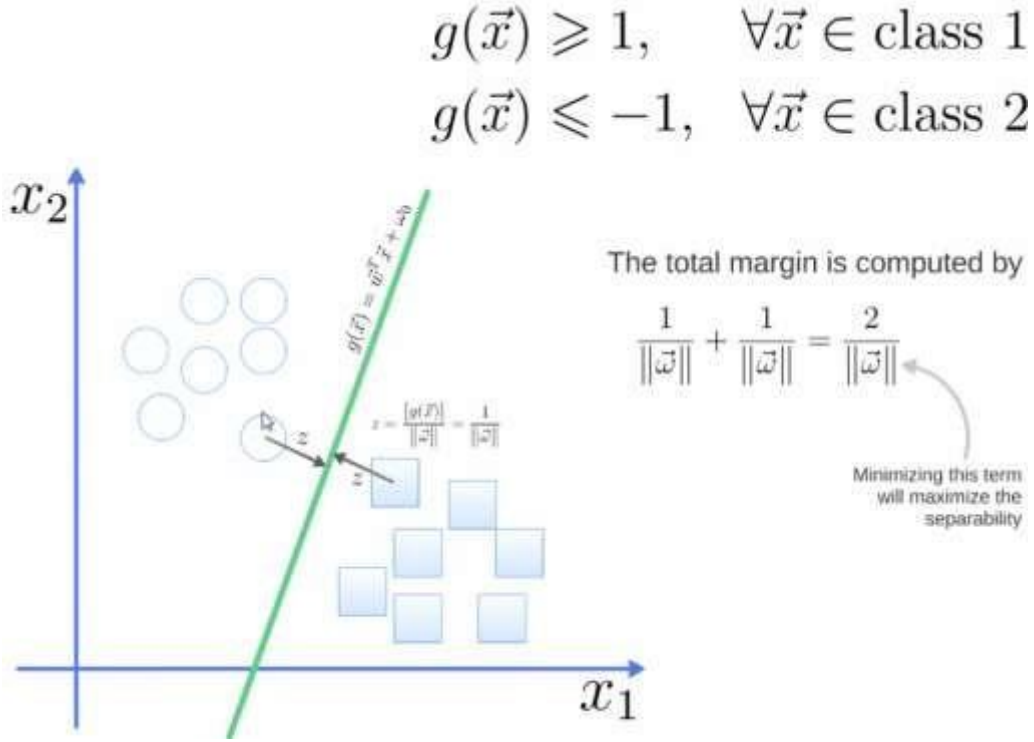
Lúc này ta cần giải toán tối ưu:

$$Min\left(L(w)\right) = \frac{1}{2}\|w\|^2$$

$$y_i(x_i.w^T + b) \geq 1, i = 1, .., l$$

ĐK Karush-Kuhn-Tucker, sử dụng:

$$\vec{w} = \sum_{i=0}^{N} \lambda_i y_i x_i \qquad \sum_{i=0}^{N} \lambda_i y_i = 0$$



Ví dụ

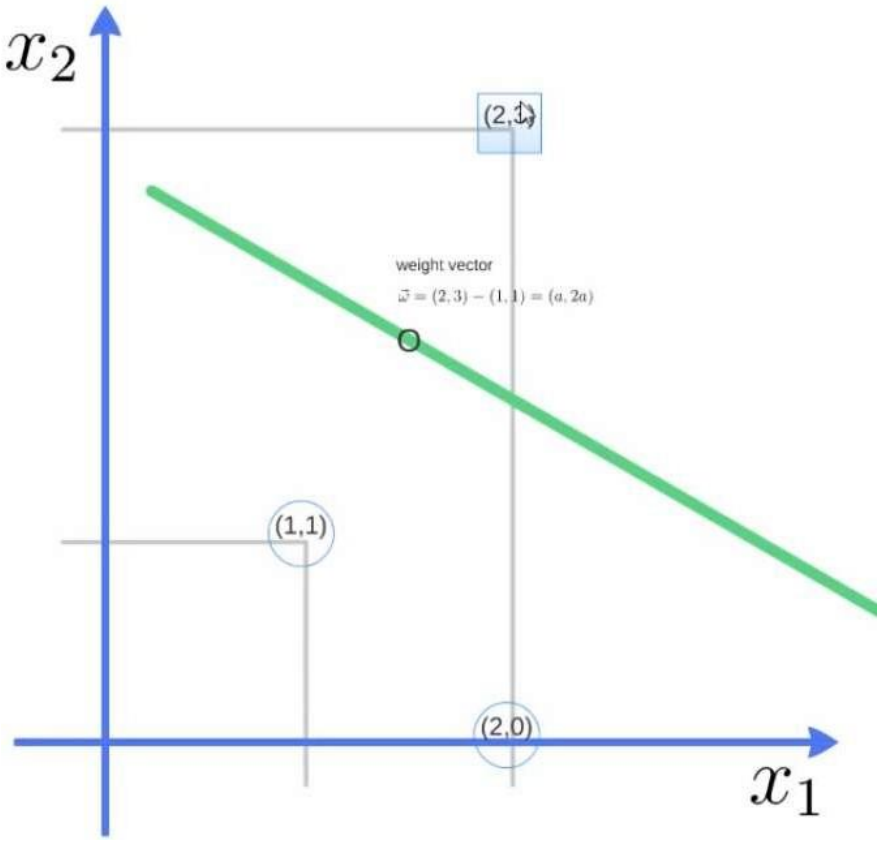
Ta có: $g(x) = \vec{w}^T . x + w_0$,

$\vec{w} = (a, 2a)$

Từ đó suy ra: a = 2/5, w_0 = -11/5

$\vec{w} = \left(\frac{2}{5}, \frac{4}{5}\right)$

$g(x) = x_1 + 2x_2 - 5{,}5$



4. Bài toán phân 2 lớp với SVM

TH2: Tập D phân chia tuyến tính có nhiễu

Đặt $f(\vec{x_i}) = \text{sign}(\vec{x_i}.\vec{w} + b) = \begin{cases} +1, \vec{x_i}\vec{w} + b > 0 \\ -1, \vec{x_i}.\vec{w} + b < 0 \end{cases}$

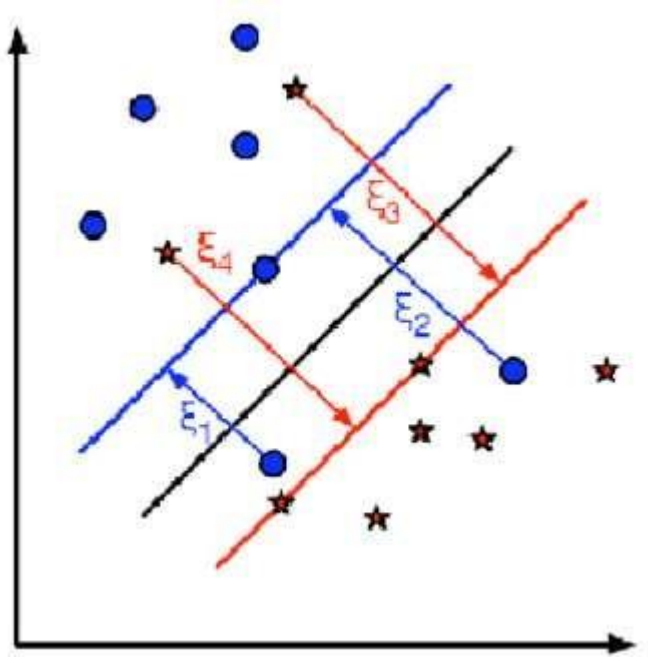
Sử dụng $\varepsilon_i \geq 0$: $y_i.(x_i.\vec{w} + b) \geq 1 - \varepsilon_i$, i=1,…,l

Lúc này ta cần giải toán tối ưu:

$$Min\left(L(w,\varepsilon)\right) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\varepsilon_i$$

$$y_i(x_i.w^T + b) \geq 1 - \varepsilon_i, i = 1,..,l; \varepsilon_i \geq 0$$

Trong đó: C là tham số cho trước



4. Bài toán phân 2 lớp với SVM

TH3: Tập D không phân chia tuyến tính

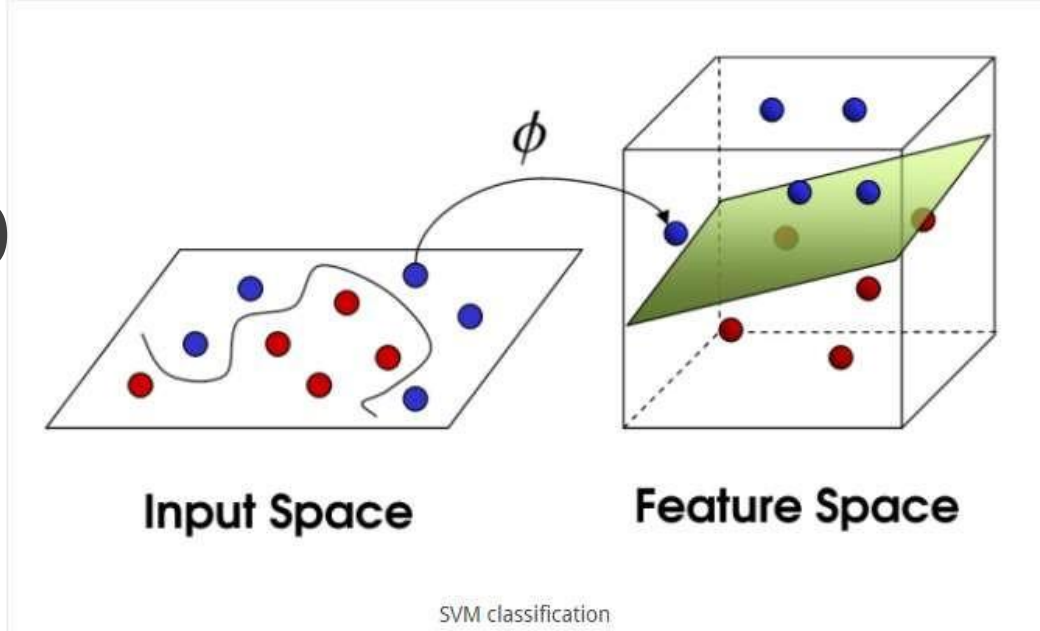
Gọi Ø là ánh xạ phi tuyến từ không gian R^n vào không gian R^m

$$Ø: R^n \rightarrow R^m$$

Lúc này ta cần giải toán tối ưu:

$$Min\left(L(w,\varepsilon)\right) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{l}\varepsilon_i$$
$$y_i\left(Ø(x_i).w^T + b\right) \geq 1 - \varepsilon_i, i = 1,..,l; \varepsilon_i \geq 0$$

Trong đó: C là tham số cho trước



SVM classification

Các bước chính của SVM

- Tiền xử lý dữ liệu: Vector của các số thực (Nếu chưa phải là số thực thì chuyển về dạng số SVM, tránh các số quá lớn, thường co giãn dữ liệu [-1,1] hoặc [0,1])
- Chọn hàm hạt nhân: phù hợp cho từng bài toán cụ thể để được độ chính xác cao
- Thực hiện việc kiểm tra chéo để xác định các tham số cho ứng dụng
- Sử dụng các tham số cho việc huấn luyện tập mẫu
- Kiểm thử tập dữ liệu Test

So sánh và một số cải tiến

So sánh

- Không cần xác định mô hình của đối tượng như neuron, fuzy logic, mạng fuzzy-neuron,…
- Nó luôn kết hợp với các dữ liệu có ý nghĩa về mặt vật lý, do đó dễ dàng giải thích được tường minh.

Cải tiến

- Một số biến thể như C-SVC, V-SVC
- Thuật toán NNSRM (Nearest Neighbor Structural Risk Minimization): kết hợp giữa 2 kỹ thuật SVM và Nearest Neighbor.

5. Đánh giá và Kết Luận

➢ SVM+ Kernel Methods

- ❑ Phương pháp học mới.
- ❑ Cung cấp nhiều công cụ.
- ❑ Nền tảng lý thuyết học thống kê.
- ❑ Tối ưu toàn cục, mô hình chất lượng cao, chịu đựng được nhiễu.
- ❑ Thành công trong nhiều ứng dụng.

➢ Hạn chế

- ❑ Khó dịch kết quả
- ❑ Độ phức tạp vẫn cao
- ❑ Xử lý dữ liệu kiểu số
- ❑ Tham số đầu vào

5. Đánh giá và Kết Luận

Kết luận

- Cơ sở lý thuyết của phương pháp SVM
- Phương pháp phân lớp hiệu quả được nghiên cứu nhiều nhất trong thời gian qua
- Kết hợp với một số phương pháp khác để tăng hơn nữa tốc độ tính toán, cũng như độ chính xác cho SVM
- Cải tiến cho phép phân chia không gian dữ liệu một cách tốt hơn, nhằm loại bỏ những vùng không được phân lớp bằng cách đưa kỹ thuật mờ vào SVM.

5. Đánh giá và Kết Luận

Hướng phát triển

- Multi-class
- Clustering
- Xử lý dữ liệu lớn
- Dữ liệu không phải kiểu số
- Dữ liệu không cân bằng
- Xây dựng hàm nhân
- Dịch kết quả
- Tìm kiếm thông tin (ranking)