KHAI PHÁ DỮ LIỆU

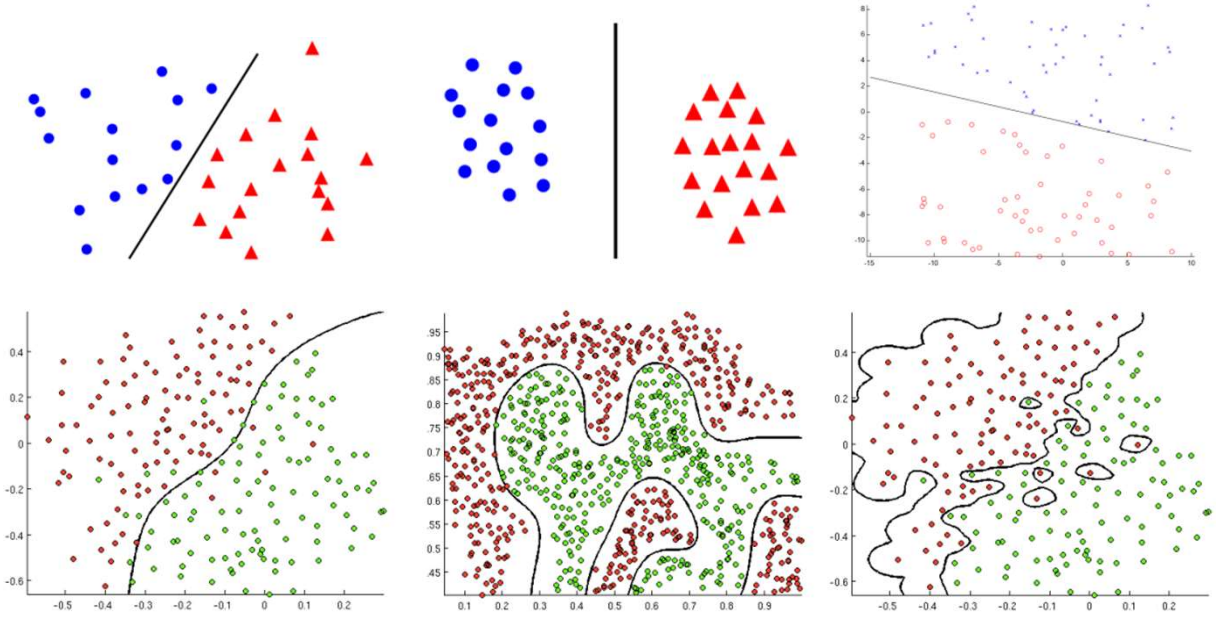
Trường Đại học Nha Trang
Khoa Công nghệ thông tin
Bộ môn Hệ thống thông tin
Giáo viên: TS.Nguyễn Khắc Cường

CHỦ ĐỀ 4

PHÂN LỚP (SVM)

SVM

- SVM = Support Vector Machines
- Là một trong các Supervised learning methods
- Dùng được để thực hiện classification
  - Binary classification
  - Multi-class classification
- Ví dụ:

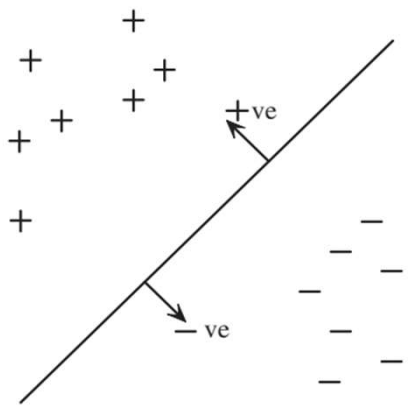


SVM

- Idea toán học thực hiện Binary classification:
  - Biến đổi mỗi sample trong dataset thành một vector

$$\vec{x} \in R^n \quad (n = 1...\infty)$$

  - Hai class được mã hóa thành
    - Class +1
    - Class -1



  - Mỗi sample được gán nhãn: $(\vec{x}, y)$ $y \in \{-1, +1\}$

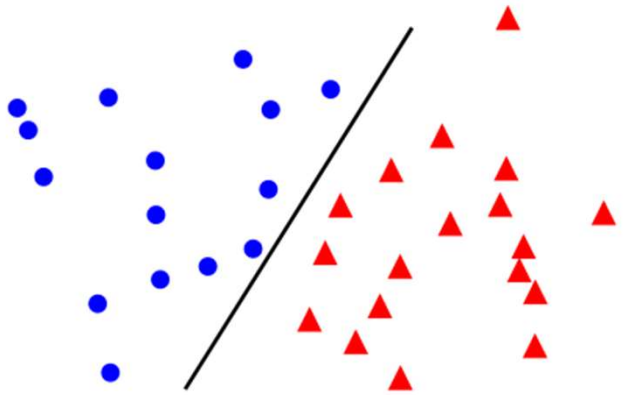
SVM

- Idea toán học thực hiện Binary classification:
  - Thu được training dataset: $(\vec{x}_i, y_i) \qquad i = 1...N$
  - Xây dựng hyperplane ngăn cách 2 lớp, có dạng:

$$f(\vec{x}) = \vec{w}^T \vec{x} + \vec{b}$$

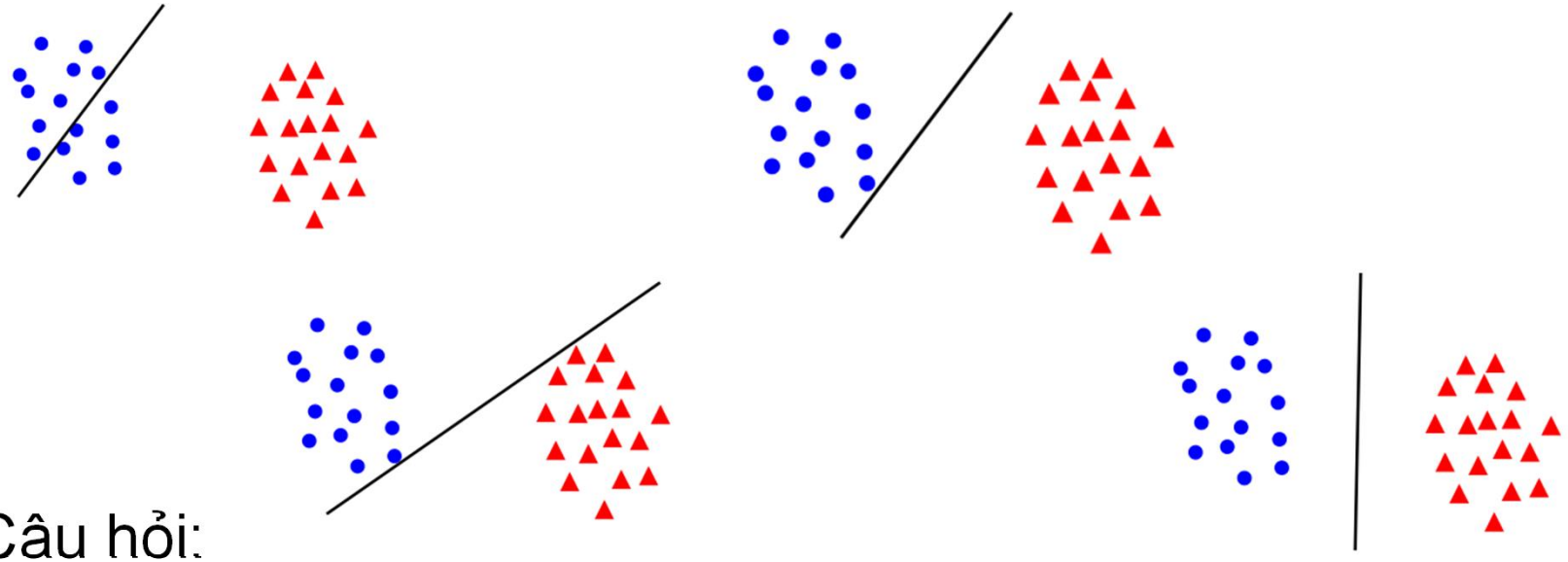
trong đó: $\vec{w}$ là normal (hay weight) vector của hyperplane
và $\vec{b}$ là bias

- Ví dụ: hyperplane trong R^2



SVM

- Idea toán học thực hiện Binary classification:
  - Nếu tập hợp các data cần phân lớp là một tập hợp có thể **phân chia một cách tuyến tính** được, thì các trường hợp có thể có của các hyperplane là:



  - Câu hỏi:
    - hyperplane nào là tốt nhất trong việc phân chia 2 class?

SVM

- Idea toán học thực hiện Binary classification:
  - Tìm các hyperplane phân lớp đúng?
    - Xét các hyperplane có thể có

$$f(\vec{x}) = \vec{w}^T \vec{x} + \vec{b}$$

    - Dựa vào tập training set $(\vec{x}_i, y_i) \quad i = 1...N,$ các hyperplane phân lớp đúng sẽ là các hyperplane $f()$ thỏa điều kiện sau:

$$f(\vec{x}_i) \begin{cases} \geq 0 & y_i = +1 \\ < 0 & y_i = -1 \end{cases}$$

    - Hay: mỗi hyperplane được xem là phân lớp đúng đối các training data nếu:

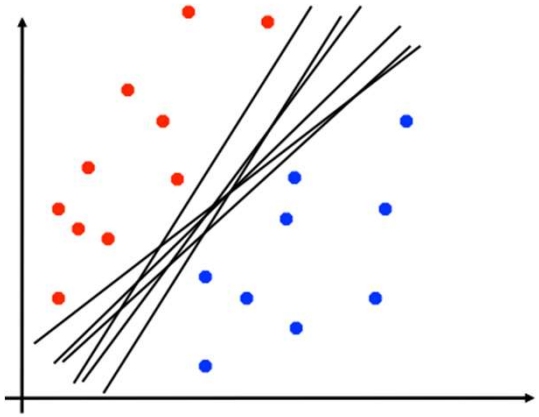
$$y_i f(\vec{x}_i) > 0 \qquad i = 1...N$$

hay:

$$y_i(\vec{w}_i^T \vec{x}_i + \vec{b}_i) > 0 \qquad i = 1...N$$

SVM

- Idea toán học thực hiện Binary classification:
  - Tìm các hyperplane phân lớp đúng?
    - Các hyperplane phân lớp đúng gồm có



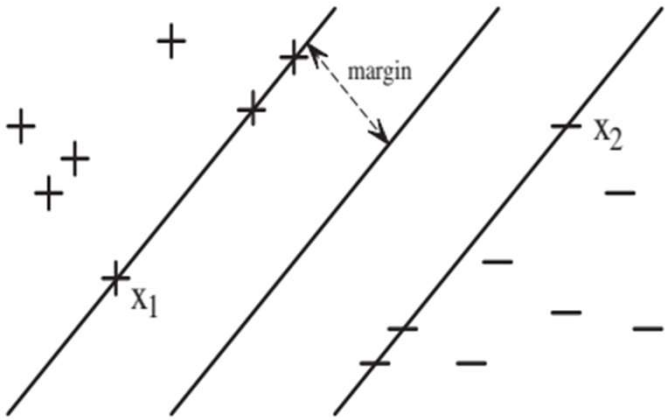
  - Câu hỏi:
    - Trong số các hyperplane phân lớp đúng đối với các data trong training dataset thì chọn hyperplane nào là tốt nhất?

SVM


V. Vapnik

- Idea toán học thực hiện Binary classification:
  - Trả lời:
    - Vapnik đã đề xuất idea SVM (1990's) để tìm ra hyperplane đó
    - Hyperplane cần tìm có largest margin
  - Margin?

$$d(\vec{x}_i) = \frac{\left|\vec{x}_i \cdot \vec{w}_i + \vec{b}\right|}{\left\|\vec{w}_i\right\|_2} = \frac{\left|\vec{x}_i \cdot \vec{w}_i + \vec{b}\right|}{\sqrt{\sum_{i=1}^{d} w_i^2}}$$



    - Margin = khoảng cách giữa hyperplane và data point gần nhất (support vectors)

SVM

- Idea toán học thực hiện Binary classification:
  - Largest margin?
    - Xét hyperplane $\vec{w}^T \vec{x} + \vec{b} = 0$
    - Thì $c(\vec{w}^T \vec{x} + \vec{b}) = 0$ cũng chính là hyperplane đó
    - Do đó, chọn normal vector w sao cho:

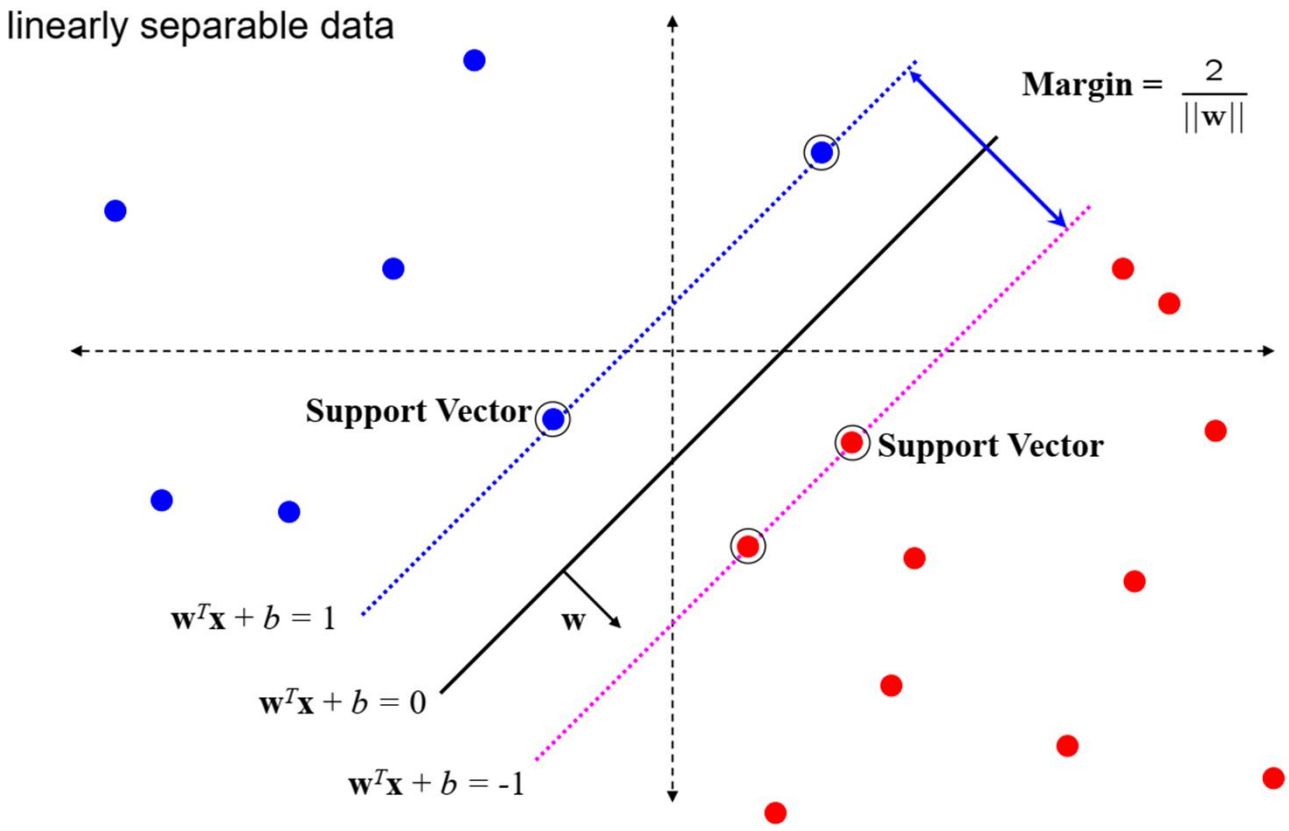
      $\vec{w}^T \vec{x}_+ + \vec{b} = +1$ và $\vec{w}^T \vec{x}_- + \vec{b} = -1$

      tương ứng với các support vectors + và –
    - Như vậy, margin tính theo w đã chọn trên là:

$$\frac{\vec{w}^T (\vec{x}_+ - \vec{x}_-)}{\|w\|} = \frac{2}{\|w\|}$$

SVM

- Idea toán học thực hiện Binary classification:
  - Largest margin?
    - Như vậy, margin tính theo w đã chọn trên là:



SVM

- Idea toán học thực hiện Binary classification:
  - Largest margin?
    - Như vậy, largest margin có thể được tìm thấy nhờ bài toán tối ưu

$$\max_{w} \frac{2}{\|w\|} \; subject \; to \; \vec{w}^T \vec{x}_i + \vec{b} \begin{cases} \geq 1 \;\; if \;\; y_i = +1 \\ \leq 1 \;\; if \;\; y_i = -1 \end{cases} (i = 1...N)$$

    - Nhận xét:
      - Đây là bài toán quadratic optimization thỏa ràng buộc tuyến tính
      - Bài toán này có nghiệm duy nhất

      → Kết quả là tìm được normal vector w

      → biểu diễn hyperplane duy nhất có largest margin đối với training dataset đã cho (có thể phân chia tuyến tính)

Q / A