LOGO

LẬP TRÌNH CHO KHOA HỌC DỮ LIỆU

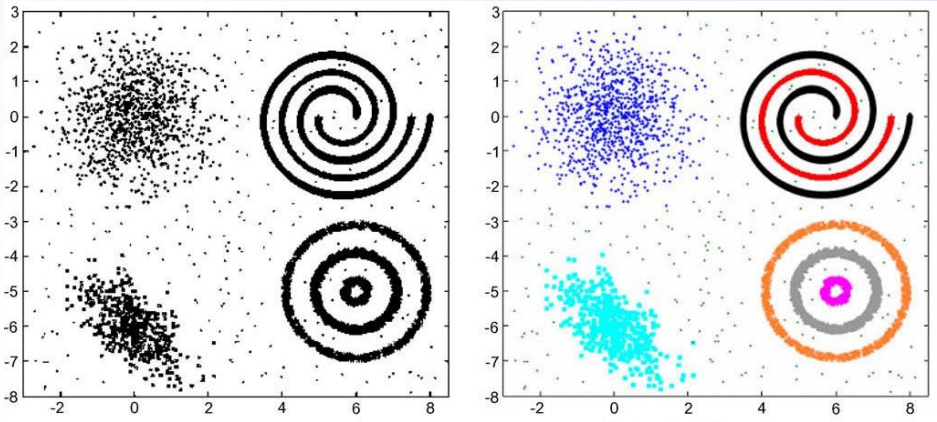
Bài 11. Một số mô hình học máy

Nội dung

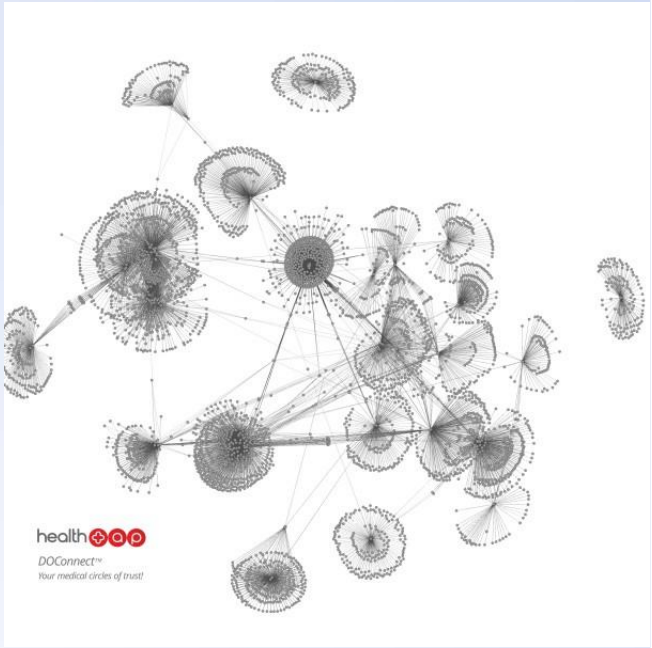
1. Phân cụm dữ liệu
2. Phân cụm mờ
3. Hồi quy tuyến tính
4. Phân lớp SVM

Phân cụm K-mean

- Phân cụm (clustering)
  - Phát hiện các cụm dữ liệu, cụm tính chất,…

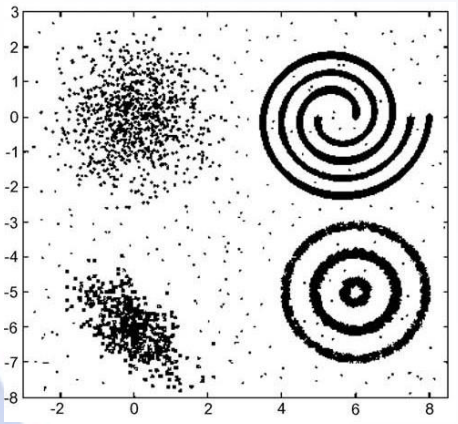


- Community detection
  - Phát hiện các cộng đồng trong mạng xã hội

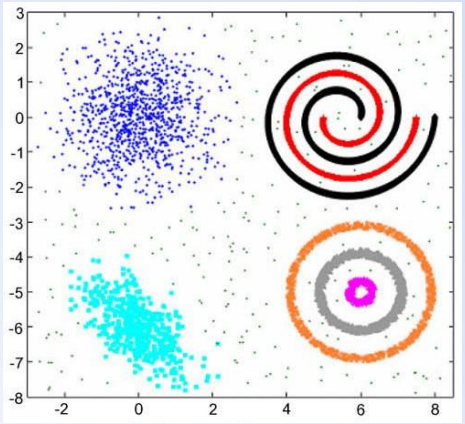


Phân cụm K-mean

- Phân cụm (clustering)
  - Đầu vào: một tập dữ liệu không có nhãn (các ví dụ không có nhãn lớp hoặc giá trị đầu ra mong muốn)
  - Đầu ra: các cụm (nhóm) của các ví dụ
- Một **cụm (cluster)** là một tập các ví dụ
  - Tương tự với nhau (theo một ý nghĩa, đánh giá nào đó)
  - Khác biệt với các ví dụ thuộc các cụm khác

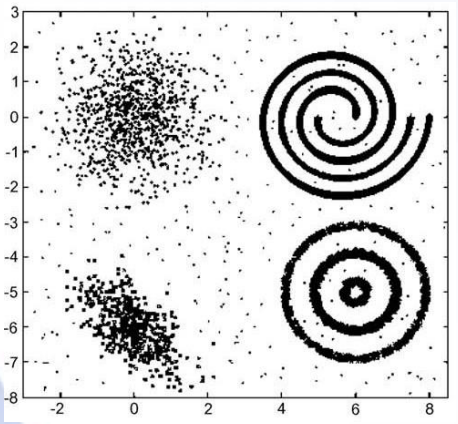


Sau khi phân cụm

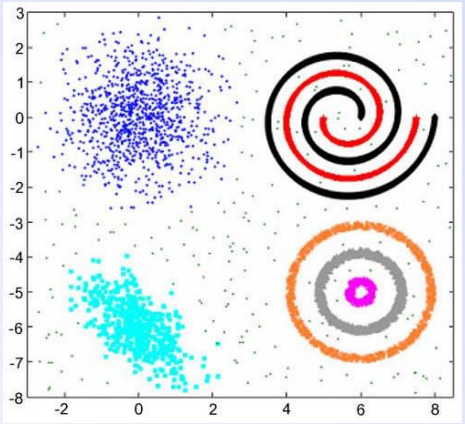


Phân cụm K-mean

- Phân cụm (clustering)
  - Đầu vào: một tập dữ liệu không có nhãn (các ví dụ không có nhãn lớp hoặc giá trị đầu ra mong muốn)
  - Đầu ra: các cụm (nhóm) của các ví dụ
- Một **cụm (cluster)** là một tập các ví dụ
  - Tương tự với nhau (theo một ý nghĩa, đánh giá nào đó)
  - Khác biệt với các ví dụ thuộc các cụm khác



Sau khi phân cụm



Phân cụm K-mean

- Giải thuật phân cụm
  - **Dựa trên phân hoạch (Partition-based clustering)**
  - **Dựa trên tích tụ phân cấp (Hierarchical clustering)**
  - Bản đồ tự tổ thức (Self-organizing map – SOM)
  - Các mô hình hỗn hợp (Mixture models)
  - …
- Đánh giá chất lượng phân cụm (Clustering quality)
  - Khoảng cách/sự khác biệt *giữa các cụm* → Cần được *cực đại hóa*
  - Khoảng cách/sự khác biệt *bên trong một cụm* → Cần được *cực tiểu hóa*

Phân cụm K-mean

- K-means được giới thiệu đầu tiên bởi Lloyd năm 1957.
- Là phương pháp phân cụm phổ biến nhất trong các phương pháp dựa trên phân hoạch (partition-based clustering)
- Biểu diễn dữ liệu: $D=\{x_1,x_2,\ldots,x_r\}$
  - x_i là một ví dụ (một vectơ trong một không gian *n* chiều)
- Giải thuật K-means phân chia tập dữ liệu thành k cụm
  - Mỗi cụm (cluster) có một điểm trung tâm, được gọi là **centroid**
  - k (tổng số các cụm thu được) là một giá trị được cho trước (vd: được chỉ định bởi người thiết kế hệ thống phân cụm)

Phân cụm K-mean

Đầu vào: tập học **D**, số lượng cụm k, khoảng cách d(x,y)

- Bước 1. Chọn ngẫu nhiên k ví dụ (được gọi là **các hạt nhân – seeds**) để sử dụng làm *các điểm trung tâm ban đầu* (*initial centroids*) của *k* cụm.
- Bước 2. Lặp liên tục hai bước sau cho đến khi *gặp điều kiện hội tụ* (convergence criterion):
  - Bước 2.1. Đối với mỗi ví dụ, *gán nó vào cụm* (trong số k cụm) mà có tâm (centroid) gần ví dụ đó nhất.
  - Bước 2.2. Đối với mỗi cụm, *tính toán lại điểm trung tâm (centroid) của nó* dựa trên tất cả các ví dụ thuộc vào cụm đó.

Phân cụm K-mean

Input

- Cho n điểm. Mỗi điểm có dạng (x, y)
- k: số nhóm. k <= n

Output

- Danh sách k nhóm và các điểm của mỗi nhóm

Flow

- B1: Đầu tiên chọn random k điểm trong tập n điểm kia làm trọng tâm của k nhóm
- B2: Tính khoảng cách từ mỗi điểm đến trọng tâm `(cx, cy)` của từng nhóm Sử dụng công thức tính khoảng cách:

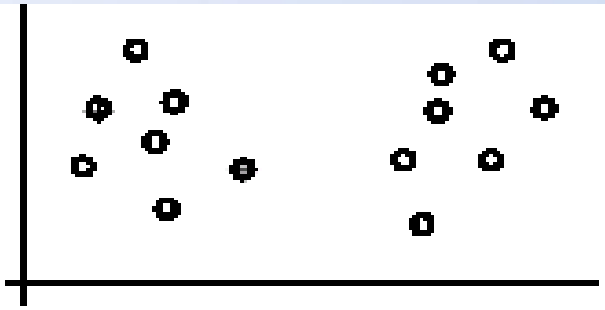
```
d(a, b) ^ 2 = (ax - bx) ^ 2 + (ay - by) ^ 2
```

- B3: Đối với mỗi điểm A, cho vào nhóm có khoảng cách từ trọng tâm nhóm đó đến điểm A là gần nhất.
- B4: Tính lại tọa độ trọng tâm cho mỗi nhóm

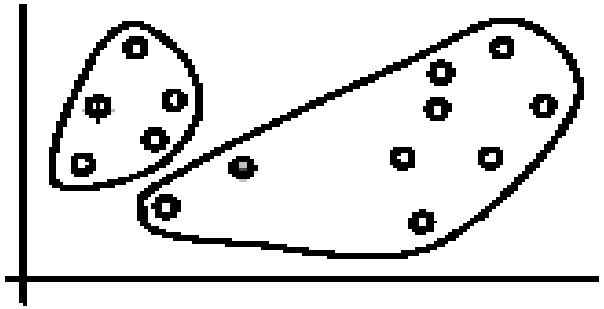
```
cx = SUM(a1.x + a2.x + ... + am.x) / m
cy = SUM(a1.y + a2.y + ... + am.y) / m

m: số phần từ của nhóm
```

- B5: Nếu `cx, cy` không đổi → đó là giá trị trọng tâm cần tìm. Else làm lại từ B2.

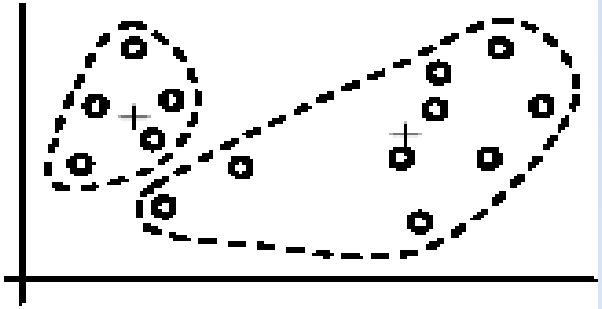
Phân cụm K-mean



(A). Random selection of k centers



Iteration 1: (B). Cluster assignment



(C). Re-compute centroids

Phân cụm K-mean

- Mặc dù có những nhược điểm như trên, k-means vẫn là giải thuật phổ biến nhất được dùng để giải quyết các bài toán phân cụm – do tính đơn giản và hiệu quả.
  - Các giải thuật phân cụm khác cũng có các nhược điểm riêng.
- Về tổng quát, không có lý thuyết nào chứng minh rằng một giải thuật phân cụm khác hiệu quả hơn k-means.
  - Một số giải thuật phân cụm có thể phù hợp hơn một số giải thuật khác đối với một số kiểu tập dữ liệu nhất định, hoặc đối với một số bài toán ứng dụng nhất định.
- So sánh hiệu năng của các giải thuật phân cụm là một nhiệm vụ khó khăn (thách thức).
  - Làm sao để biết được các cụm kết quả thu được là chính xác?

Phân cụm FCM

Phương pháp phân cụm

- Phân cụm rõ: dữ liệu được chia vào các cụm, trong đó mỗi điểm dữ liệu thuộc vào chính xác một cụm.
- Phân cụm mờ: các điểm dữ liệu có thể thuộc vào nhiều hơn một cụm và tương ứng với các điểm dữ liệu là ma trận độ thuộc.
- Phân cụm mờ bán giám sát: là phân cụm mờ kết hợp với các thông tin bổ trợ hình thành lên nhóm các thuật toán gọi là phân cụm mờ bán giám sát.

Phân cụm FCM

❖ Thuật toán Fuzzy C-means

- **Hàm mục tiêu**

$$J = \sum_{k=1}^{N}\sum_{j=1}^{C} u_{kj}^{m}\left\|X_k - V_j\right\|^2 \rightarrow \min$$

- **Điều kiện ràng buộc**

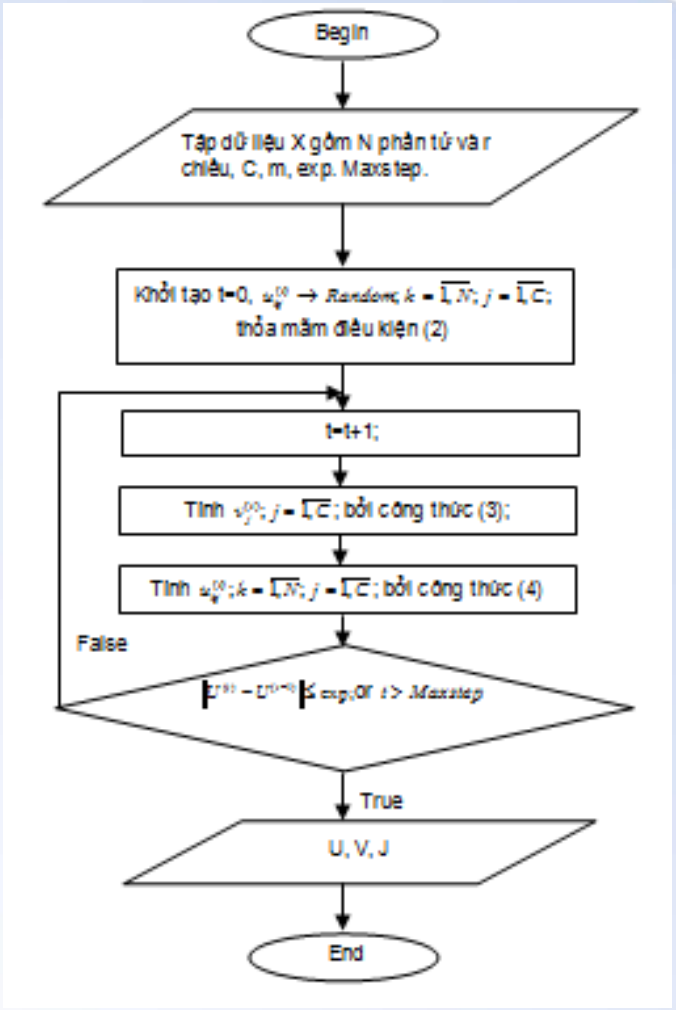
$$\sum_{j=1}^{C} u_{kj} = 1; \qquad u_{kj} \in [0,1]; \qquad \forall k = \overline{1,N}$$

- **Tính tâm cụm**

$$V_j = \frac{\sum_{k=1}^{C} u_{kj}^{m} X_k}{\sum_{k=1}^{C} u_{kj}^{m}}$$

- **Tính hàm mức độ thành viên**

$$u_{kj} = \frac{1}{\sum_{i=1}^{C}\left(\frac{\left\|X_k - V_j\right\|}{\left\|X_k - V_i\right\|}\right)^{\frac{1}{m-1}}}$$



Hồi quy tuyến tính

- Hồ quy tuyến tính: DL được mô hình hóa phù hợp với 1 đường thẳng
  - Thường dùng phương pháp bình phương tối thiểu để khớp với đường
- Hồ quy đa chiều: Cho một biến đích Y được mô hình hóa như ột hàm tuyến tính của vector đặc trưng đa chiều
- Mô hình tuyến tính loga: rời rạc hóa xấp xỉ các phân bố xác suất đa chiều

Hồi quy tuyến tính

- **Hồi quy tuyến tính**: $Y = \alpha + \beta X$
  - Hai tham số, α và β đặc trưng cho đường và được xấp xỉ qua dữ liệu đã nắm bắt được.
  - Sử dụng chiến lược BP tối thiểu tới các giá trị đã biết $Y_1, Y_2, \ldots, X_1, X_2, \ldots$
- **Hồi quy đa chiều**: $Y = b0 + b1\ X1 + b2\ X2.$
  - Nhiều hàm không tuyến tính được chuyển dạng như trên.
- **Mô hình tuyến tính loga**:
  - Bảng đa chiều của xác suất tích nối được xấp xỉ bởi tích của các bảng bậc thấp hơn
  - Xác suất: $p(a, b, c, d) = \alpha ab\ \beta ac\ \chi ad\ \delta bcd$

Phân lớp SVM

- Máy vectơ hỗ trợ (Support vector machine - SVM) được đề cử bởi V. Vapnik và các đồng nghiệp của ông vào những năm 1970s ở Nga, và sau đó đã trở nên nổi tiếng và phổ biến vào những năm 1990s
- SVM là một phương pháp **phân lớp tuyến tính** (linear classifier), với mục đích xác định một siêu phẳng (hyperplane) để phân tách **hai lớp** của dữ liệu.
  Ví dụ: lớp có nhãn dương (positive) và lớp có nhãn âm (negative)
- **Các hàm nhân (kernel functions)**, cũng được gọi là các hàm biến đổi (transformation functions), được dùng cho các trường hợp phân lớp phi tuyến

Phân lớp SVM

- SVM có một nền tảng lý thuyết chặt chẽ
- SVM là một phương pháp tốt (phù hợp) đối với những bài toán phân lớp có không gian rất nhiều chiều (các đối tượng cần phân lớp được biểu diễn bởi một tập rất lớn các thuộc tính)
- SVM đã được biết đến là một trong số các phương pháp phân lớp tốt nhất đối với các bài toán phân lớp văn bản (text classification)

Phân lớp SVM

- *Các vectơ được ký hiệu bởi các chữ đậm nét!*
- Biểu diễn tập r các ví dụ huấn luyện (training examples)

$$\{(\mathbf{x_1}, y_1), (\mathbf{x_2}, y_2), \ldots, (\mathbf{x_r}, y_r)\},$$

  - $\mathbf{x_i}$ là một **vectơ đầu vào** được biểu diễn trong không gian $X \subseteq R^n$
  - y_i là một **nhãn lớp** (giá trị đầu ra), $y_i \in \{1,-1\}$
  - y_i=1: lớp *dương* (positive); y_i=-1: lớp *âm* (negative)

- SVM xác định một hàm phân tách tuyến tính

$$f(\mathbf{x}) = \langle \mathbf{w} \cdot \mathbf{x} \rangle + b$$

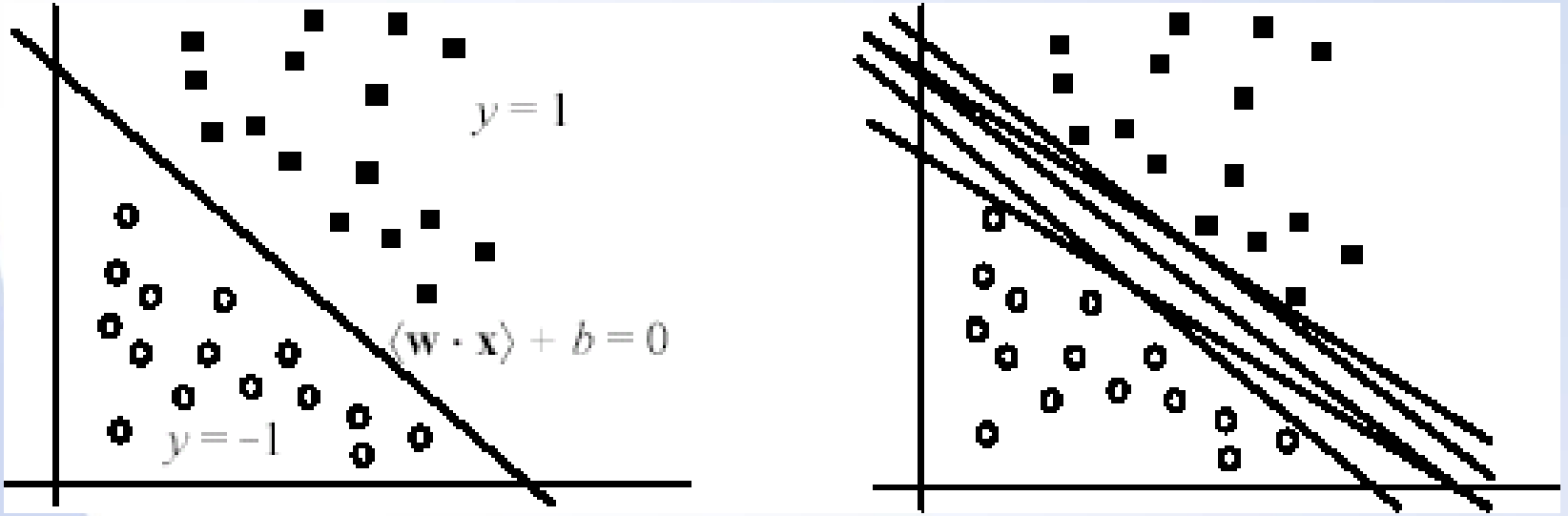
  - **w** là vectơ trọng số các thuộc tính; b là một giá trị số thực

- Sao cho với mỗi $\mathbf{x_i}$:

$$y_i = \begin{cases} 1 & if \quad \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 0 \\ -1 & if \quad \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b < 0 \end{cases}$$

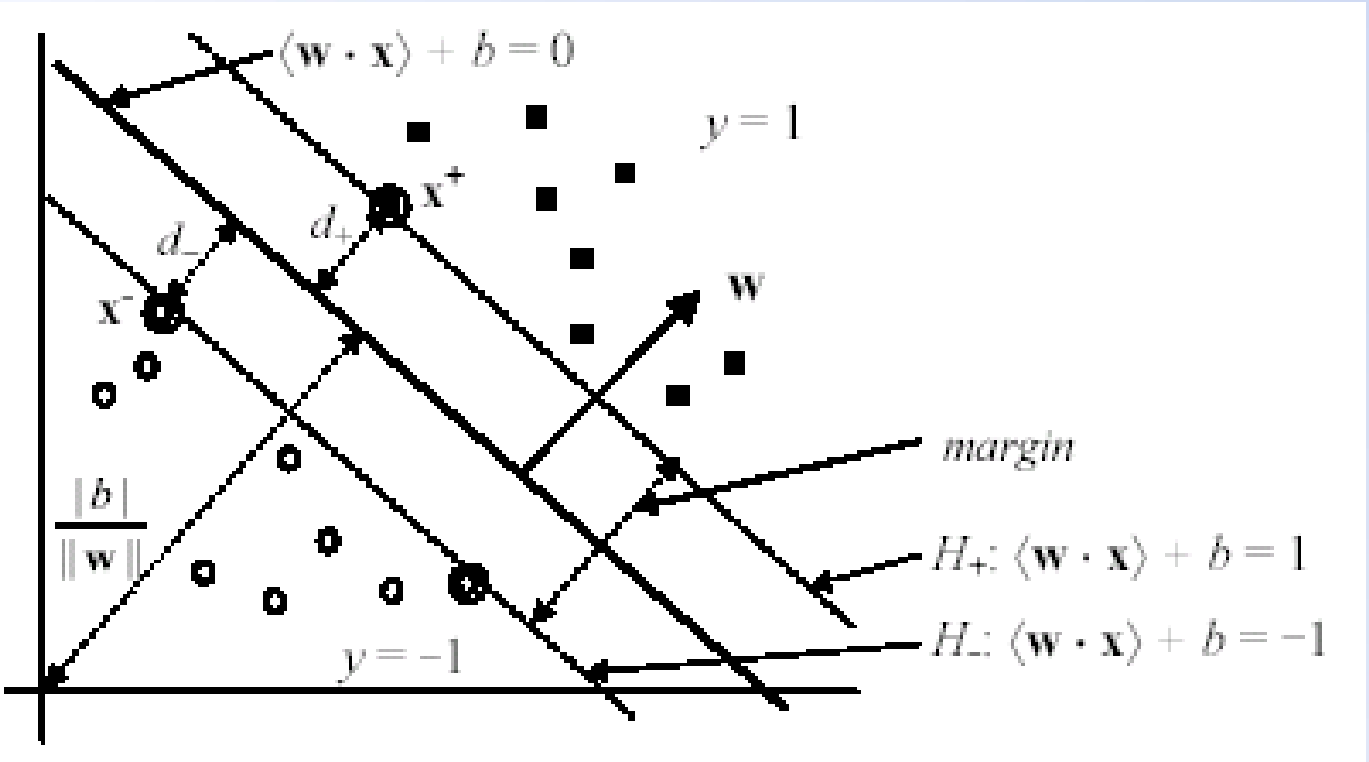
Phân lớp SVM

- Siêu phẳng phân tách các ví dụ huấn luyện lớp dương và các ví dụ huấn luyện lớp âm: $\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$
- Còn được gọi là ranh giới (bề mặt) quyết định
- Tồn tại nhiều siêu phẳng phân tách. Chọn cái nào?



Phân lớp SVM

- SVM lựa chọn mặt siêu phẳng phân tách có lề (margin) lớn nhất
- Lý thuyết học máy đã chỉ ra rằng *một mặt siêu phẳng phân tách như thế sẽ tối thiểu hóa giới hạn lỗi (phân lớp) mắc phải (so với mọi siêu phẳng khác)*



Phân lớp SVM

- Giả sử rằng tập dữ liệu (tập các ví dụ huấn luyện) có thể phân tách được một cách tuyến tính
- Xét một ví dụ của lớp dương ($\mathbf{x}^+$,1) và một ví dụ của lớp âm ($\mathbf{x}^-$,-1) *gần nhất* đối với siêu phẳng phân tách H_0 ($\langle \mathbf{w}\cdot\mathbf{x}\rangle + b = 0$)
- Định nghĩa 2 siêu phẳng lề song song với nhau
  - H_+ đi qua $\mathbf{x}^+$, và song song với H_0
  - H_- đi qua $\mathbf{x}^-$, và song song với H_0

H_+: $\langle \mathbf{w}\cdot\mathbf{x}^+\rangle + b = 1$

H_-: $\langle \mathbf{w}\cdot\mathbf{x}^-\rangle + b = -1$ [Eq.3]

sao cho: $\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b \geq 1$, nếu $y_i = 1$

$\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + b \leq -1$, nếu $y_i = -1$

Phân lớp SVM

- **Mức lề** (margin) là khoảng cách giữa 2 siêu phẳng lề H_+ và H_-. Trong hình vẽ nêu trên:
  - d_+ là khoảng cách giữa H_+ và H_0
  - d_- là khoảng cách giữa H_- và H_0
  - $(d_+ + d_-)$ là mức lề
- Trong không gian vectơ, **khoảng cách** từ một điểm $\mathbf{x}_i$ đến siêu phẳng $(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0)$ là:

$$\frac{|\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|}$$

trong đó $\|\mathbf{w}\|$ là độ dài của $\mathbf{w}$:

$$\|\mathbf{w}\| = \sqrt{< \mathbf{w} \cdot \mathbf{w} >} = \sqrt{w_1^2 + w_1^2 + \ldots + w_1^2}$$

Phân lớp SVM

- Tính toán d_+: khoảng cách từ $\mathbf{x}^+$ đến ($\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$)
  - Áp dụng các biểu thức [Eq.3-4]:

$$d_+ = \frac{|\langle \mathbf{w} \cdot \mathbf{x}^+ \rangle + b|}{\|\mathbf{w}\|} = \frac{|1|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

- Tính toán d_-: khoảng cách từ $\mathbf{x}^-$ đến ($\langle \mathbf{w} \cdot \mathbf{x} \rangle + b = 0$)
  - Áp dụng các biểu thức [Eq.3-4]:

$$d_- = \frac{|\langle \mathbf{w} \cdot \mathbf{x}^- \rangle + b|}{\|\mathbf{w}\|} = \frac{|-1|}{\|\mathbf{w}\|} = \frac{1}{\|\mathbf{w}\|}$$

- Tính toán mức lề

$$margin = d_+ + d_- = \frac{2}{\|\mathbf{w}\|}$$

Phân lớp SVM

Định nghĩa (**Linear SVM –** Trường hợp **phân tách được**)

- Tập gồm r ví dụ huấn luyện có thể phân tách tuyến tính

  $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_r, y_r)\}$

- SVM học một phân lớp (classifier) mà có mức lề cực đại
- Tương đương với việc giải quyết **bài toán tối ưu bậc hai** sau đây
  - Tìm ***w*** và *b* sao cho: $margin = \frac{2}{\|\mathbf{w}\|}$ đạt cực đại
  - Với điều kiện:

$$\begin{cases} \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 1, \text{ if } y_i = 1 \\ \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \leq -1, \text{ if } y_i = -1 \end{cases}$$

với mọi ví dụ huấn luyện $\mathbf{x}_i$ (i=1..r)

Phân lớp SVM

- Học SVM tương đương với giải quyết **bài toán cực tiểu hóa có ràng buộc** sau đây

Cực tiểu hóa:

$$\frac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2}$$

Với điều kiện:

$$\begin{cases} \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 1, & if \ \ y_i = 1 \\ \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \leq -1, & if \ \ y_i = -1 \end{cases}$$

- tương đương với

Cực tiểu hóa:

$$\frac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2}$$

Với điều kiện:

$$y_i (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \quad \forall i = 1..\, r$$

Phân lớp SVM

- Học SVM tương đương với giải quyết **bài toán cực tiểu hóa có ràng buộc** sau đây

Cực tiểu hóa:

$$\frac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2}$$

Với điều kiện:

$$\begin{cases} \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \geq 1, & if \ \ y_i = 1 \\ \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b \leq -1, & if \ \ y_i = -1 \end{cases}$$

- tương đương với

Cực tiểu hóa:

$$\frac{\langle \mathbf{w} \cdot \mathbf{w} \rangle}{2}$$

Với điều kiện:

$$y_i (\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b) \geq 1, \quad \forall i = 1..\, r$$

Phân lớp SVM

- SVM chỉ làm việc với không gian đầu vào là các số thực
  - → Đối với các thuộc tính định danh (nominal), cần chuyển các giá trị định danh thành các giá trị số
- SVM chỉ làm việc (thực hiện phân lớp) với 2 lớp
  - → Đối với các bài toán phân lớp gồm nhiều lớp, cần chuyển thành một tập các bài toán phân lớp gồm 2 lớp, và sau đó giải quyết riêng rẽ từng bài toán 2 lớp này
  - → Ví dụ: chiến lược "one-against-rest"
- Siêu phẳng phân tách (ranh giới quyết định phân lớp) xác định được bởi SVM thường khó hiểu đối với người dùng
  - Vấn đề (khó giải thích quyết định phân lớp) này càng nghiêm trọng, nếu các hàm nhân (kernel functions) được sử dụng
  - SVM thường được dùng trong các bài toán ứng dụng mà trong đó việc giải thích hoạt động (quyết định) của hệ thống cho người dùng không phải là một yêu cầu quan trọng

Phân lớp SVM

Anh chị lựa chọn cho mình một trong số các mô hình sau:

- Phân lớp dựa trên Naïve bayes
- Phân lớp dựa trên SVM
- Phân cụm dữ liệu cứng K-mean
- Phân cụm dữ liệu mờ FCM
- Hồi quy tuyến tính
- Cây quyết định
- Mạng nơron


LOGO
CẢM ƠN!