BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC QUY NHƠN

HUỲNH TRỌNG PHÚC

THUẬT TOÁN SVM VỚI ÁNH XẠ ĐẶC TRƯNG TUYẾN TÍNH TỪNG PHẦN VÀ ỨNG DỤNG

ĐỀ ÁN THẠC SĨ

KHOA HỌC DỮ LIỆU

Bình Định - Năm 2025

BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC QUY NHƠN

TRẦN KHƯƠNG DUY

THUẬT TOÁN SVM VỚI ÁNH XẠ ĐẶC TRƯNG TUYẾN TÍNH TỪNG PHẦN VÀ ỨNG DỤNG

Ngành: Khoa học dữ liệu

Mã số: 8460108

Người hướng dẫn: TS. Lê Quang Thuận

Lời cam đoan

Tôi xin cam đoan nội dung trong luận văn "**Thuật toán SVM với ánh xạ đặc trưng tuyến tính từng phần và ứng dụng**" là do bản thân thực hiện theo logic riêng dưới sự hướng dẫn của TS. Lê Quang Thuận. Các nội dung và kết quả sử dụng trong luận văn đều có trích dẫn và chú thích nguồn gốc rõ ràng.

Tác giả đề án

Huỳnh Trọng Phúc

Lời cảm ơn

Lời đầu tiên tôi xin gởi đến TS. Lê Quang Thuận lời cảm ơn sâu sắc về sự tận tình giúp đỡ của thầy đối với tôi trong suốt khóa học, đặc biệt trong quá trình làm luận văn.

Tôi xin được bày tỏ lòng biết ơn đến tất cả các thầy cô Trường Đại Học Quy Nhơn, các giảng viên Trường ĐHKHTN-ĐHQG TP.HCM đã nhiệt tình giảng dạy chúng tôi trong suốt khóa học.

Xin được cảm ơn các vị lãnh đạo và chuyên viên Phòng Đào tạo sau đại học Trường Đại Học Quy Nhơn đã tạo điều kiện thuận lợi cho tôi trong suốt quá trình học.

Tôi cũng xin được cảm ơn gia đình tôi, các đồng nghiệp và các bạn học viên Cao học khóa 25B đã hỗ trợ, động viên tôi trong suốt thời gian học.

Cuối cùng, với kiến thức còn hạn chế nên dù rất cố gắng nhưng chắc chắn luận văn còn nhiều sót. Kính mong các thầy cô và các bạn đồng nghiệp đóng góp ý kiến để luận văn còn hoàn chỉnh hơn.

Trân trọng cảm ơn!

MỤC LỤC

Danh mục các bảng

Danh mục các hình

MỞ ĐẦU

Tính cấp thiết của đề tài

Ngày nay, nhiệm vụ phân loại trong học máy (machine learning) đóng một vai trò quan trọng trong nhiều ứng dụng thực tế, giúp giải quyết các vấn đề thực tiễn và cải thiện hiệu quả của nhiều quy trình. Ví dụ:

- **Ứng dụng trong y tế**: Trong y học, phân loại giúp xác định các bệnh, chẳng hạn như phân loại hình ảnh chẩn đoán (MRI, X-quang) để phát hiện các bệnh ung thư, bệnh tim mạch, hoặc các vấn đề sức khỏe khác. Sự chính xác trong phân loại có thể cứu sống bệnh nhân.

- **An ninh và bảo mật**: Trong nhận dạng khuôn mặt, nhận dạng vân tay, hay các hệ thống xác thực sinh trắc học, phân loại giúp xác định người dùng một cách nhanh chóng và chính xác. Điều này rất quan trọng trong bảo mật thông tin cá nhân và giao dịch.

- **Phân loại thư rác (Spam)**: Trong các ứng dụng email, phân loại thư điện tử thành thư rác và thư hợp lệ là một trong những nhiệm vụ phân loại quan trọng, giúp người dùng tránh bị quấy rầy bởi các thông tin không mong muốn.

- **Xử lý ngôn ngữ tự nhiên (NLP)**: Các nhiệm vụ phân loại trong NLP như phân loại cảm xúc trong văn bản, phân loại chủ đề hay phân loại văn bản theo các chủ đề cụ thể là rất quan trọng trong các ứng dụng như phân tích cảm xúc, chatbot, và các hệ thống tìm kiếm.

- **Tiếp thị và phân tích khách hàng**: Phân loại khách hàng theo các nhóm dựa trên hành vi mua sắm, sở thích hoặc mức độ gắn kết giúp các

doanh nghiệp tối ưu hóa các chiến lược marketing và gia tăng hiệu quả tiếp cận khách hàng.

- **Tự động hóa và phân tích dữ liệu lớn**: Phân loại dữ liệu là một trong những nhiệm vụ cơ bản trong phân tích dữ liệu lớn, giúp nhận diện các mẫu dữ liệu quan trọng và phân loại chúng thành các nhóm có ý nghĩa, từ đó hỗ trợ ra quyết định.
- **Quản lý rủi ro trong tài chính**: Phân loại các khoản vay là rủi ro thấp hay cao giúp các tổ chức tài chính đưa ra các quyết định cho vay, giảm thiểu rủi ro tài chính.

Các bộ phân loại tuyến tính từng phần là một mở rộng đơn giản của các bộ phân loại tuyến tính, đồng thời có khả năng xấp xỉ các ranh giới phi tuyến tính. Điều này làm cho chúng trở nên phù hợp cho các ứng dụng đòi hỏi tốc độ xử lý nhanh và ít tốn bộ nhớ, chẳng hạn như robot trinh sát nhỏ, camera thông minh, và các hệ thống nhúng.

Mục đích và nhiệm vụ nghiên cứu

Mục đích nghiên cứu

Nhiệm vụ nghiên cứu

Nhiệm vụ nghiên cứu của đề án bao gồm:

Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu

Phạm vi nghiên cứu

Cơ sở lí luận và phương pháp nghiên cứu

Cơ sở lý luận

Phương pháp nghiên cứu được áp dụng bao gồm:

Kết cấu của đề án

Ngoài phần mở đầu, kết luận, tài liệu tham khảo, phụ lục đề án có 3 chương.

- Chương 1: KIẾN THỨC CHUẨN BỊ
- Chương 2: THUẬT TOÁN SVM VỚI ÁNH XẠ ĐẶC TRƯNG TUYẾN TÍNH TỪNG PHẦN
- Chương 3: ỨNG DỤNG CỦA THUẬT TOÁN SVM VỚI ÁNH XẠ ĐẶC TRƯNG TUYẾN TÍNH TỪNG PHẦN

NỘI DUNG

CHƯƠNG 1: KIẾN THỨC CHUẨN BỊ

CHƯƠNG 2: THUẬT TOÁN SVM VỚI ÁNH XẠ ĐẶC TRƯNG TUYẾN TÍNH TỪNG PHẦN

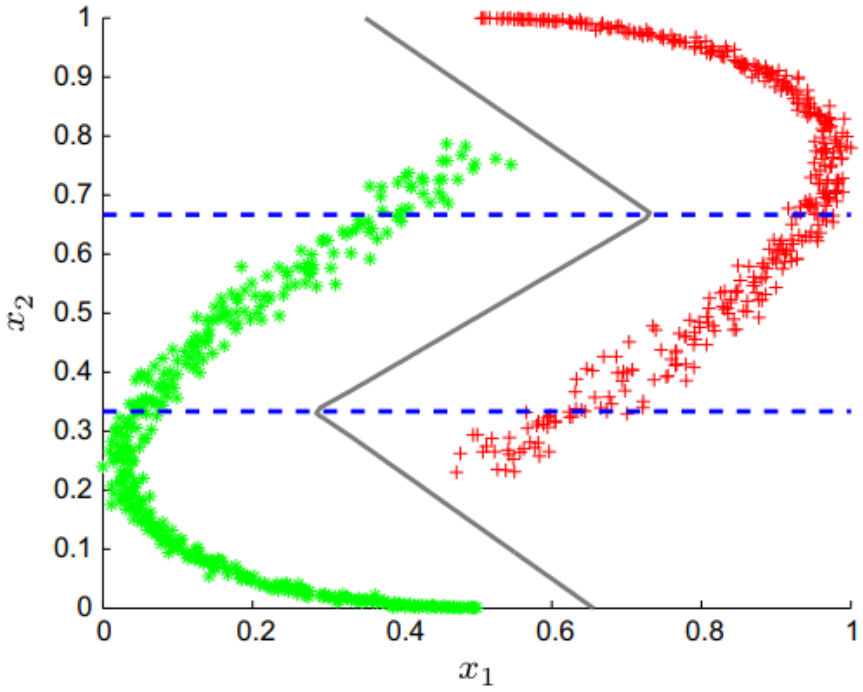
2.1. Thuật toán SVM (Support Vector Machine)

2.2. SWM với ánh xạ đặc trưng tuyến tính từng phần

2.2.1. Biên tuyến tính từng phần

Trong vấn đề phân loại hai lớp, miền thường được phân chia thành hai phần bởi một biên, ví dụ như một siêu phẳng được tạo ra bởi các phương pháp phân loại tuyến tính, theo dữ liệu đầu vào $x \in \mathbb{R}^n$ và nhãn tương ứng $y \in \{+1, -1\}$. Các bộ phân loại tuyến tính đã được nghiên cứu trong nhiều năm, tuy nhiên, khả năng phân loại của chúng quá hạn chế và cần có các bộ phân loại phi tuyến. Trong đề án này, thuật ngữ khả năng phân loại của một bộ phân loại có nghĩa là tính linh hoạt của các loại và hình dạng có thể có của biên phân loại. Ví dụ, một bộ phân loại tuyến tính chỉ có thể cung cấp một biên tuyến tính, đó là một tập hợp affine, tức là một siêu phẳng, và do đó khả năng phân loại của nó không đủ cho nhiều ứng dụng. Theo một số quan điểm, mở rộng đơn giản nhất cho một tập hợp affine là một tập hợp tuyến tính từng phần, cung cấp một biên tuyến tính từng phần. Như tên gọi, một tập hợp tuyến tính từng phần bằng với một tập hợp affine trong mỗi vùng con của miền, và các vùng con phân chia miền.

Xét tập dữ liệu hai mặt trăng được hiển thị trong Hình 1, trong đó các điểm thuộc hai lớp được đánh dấu bằng các ngôi sao màu xanh lá cây và các dấu chéo màu đỏ. Hai lớp này không thể được phân tách bởi một biên

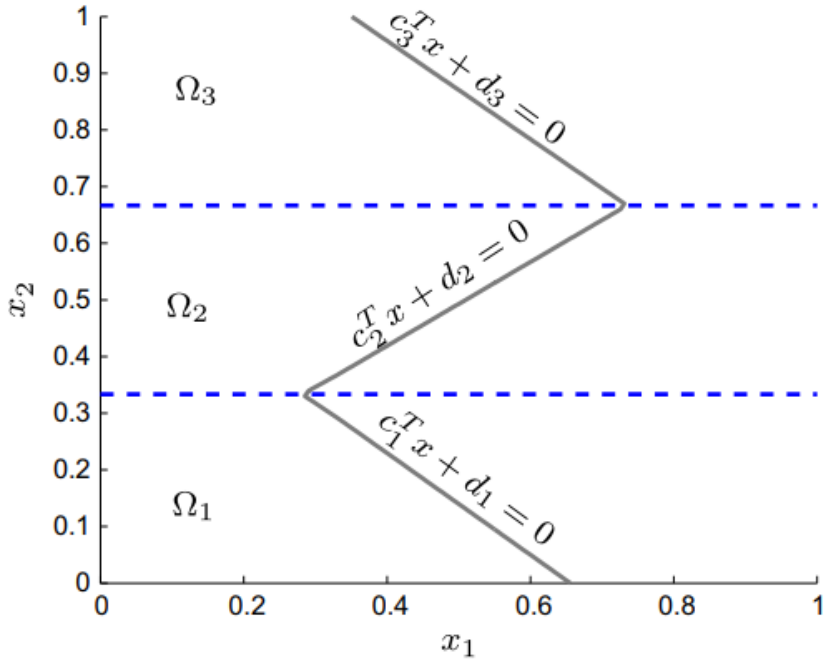


Hình 1: Tập dữ liệu hình mặt trăng. Các điểm trong hai lớp được đánh dấu bằng các ngôi sao và các dấu chéo, tương ứng; ranh giới phân loại được hiển thị bằng các đường liền nét.

tuyến tính và chúng ta có thể sử dụng một biên tuyến tính từng phần (PWL), được hiển thị bằng các đường màu đen, để phân loại hai tập dữ liệu này rất tốt. Tập hợp PWL này, được ký hiệu là $\mathcal{B}$, bao gồm ba đoạn. Mỗi đoạn có thể được định nghĩa là một đường thẳng bị giới hạn trong một vùng con. Ví dụ, chúng ta có thể phân chia miền thành $\Omega_1 = \{x : 0 \leq x(2) \leq \frac{1}{3}\}$, $\Omega_2 = \{x : \frac{1}{3} \leq x(2) \leq \frac{2}{3}\}$, $\Omega_3 = \{x : \frac{2}{3} \leq x(2) \leq 1\}$, trong đó $x(i)$ là thành phần thứ i của x. Sau đó, $\mathcal{B}$ có thể được định nghĩa là

$$\mathcal{B} = \bigcup_{k=1}^{3} \{x : c_k^T x + d_k = 0, x \in O_k\},$$

trong đó $c_k \in \mathbb{R}^2$, $d_k \in \mathbb{R}$ định nghĩa đường thẳng trong mỗi vùng con, như được hiển thị trong Hình 2.



Hình 2: Biên phân loại bị giới hạn trong mỗi vùng con, ranh giới tuyến tính từng phần tương ứng với một đường thẳng.

Để thuận tiện cho việc diễn đạt, định nghĩa của tập hợp tuyến tính từng phần được đưa ra dưới đây.

Định nghĩa 2.2.1. Nếu một tập hợp $\mathcal{B}$ được định nghĩa trong miền $\Omega \subseteq \mathbb{R}^n$ thỏa mãn hai điều kiện sau:

(i) Miền Ω được phân chia thành các đa diện hữu hạn $\Omega_1, \Omega_2, \ldots, \Omega_K$, tức là $\Omega = \bigcup_{k=1}^{K} \Omega_k$ và $\overset{\circ}{\Omega}_k \cap \overset{\circ}{\Omega}_l = \emptyset, \forall k \neq l$, trong đó $\overset{\circ}{\Omega}_k$ là phần trong của Ω_k.

(ii) Trong mỗi vùng con, $\mathcal{B}$ bằng với một tập hợp tuyến tính, tức là với mỗi k, tồn tại $c_k \in \mathbb{R}^n$, $d_k \in \mathbb{R}$ sao cho $\mathcal{B} \cap \Omega_k = \{x : c_k^T x + d_k = 0\}$.

Một tập hợp affine cung cấp một bộ phân loại tuyến tính, có thể được viết như là nghiệm của một phương trình tuyến tính $f(x) = 0$. Do đó, trong một vấn đề phân loại tuyến tính, người ta thường tìm kiếm một hàm tuyến tính

$f(x)$ để phân loại một điểm bằng dấu của giá trị hàm, tức là, sign$\{f(x)\}$. Tương tự, một tập hợp **tuyến tính từng phần (PWL)** cung cấp một biên tuyến tính từng phần và nó có thể được biểu diễn như là tập nghiệm của một phương trình tuyến tính từng phần, được đảm bảo bởi định lý sau:

Định lý 2.2.1. *Mọi tập hợp tuyến tính từng phần $\mathcal{B}$ đều có thể được biểu diễn như là nghiệm của một phương trình tuyến tính từng phần.*

Chứng minh. Theo các ký hiệu trong Định nghĩa 2.2.1, giả sử số các đa diện xác định $\mathcal{B}$ là K và các đa diện này được biểu diễn dưới dạng $\Omega_k = \{x : a_{ki}^T x + b_{ki} \leq 0, \forall 1 \leq i \leq I_k\}$, trong đó I_k là số các bất đẳng thức tuyến tính xác định Ω_k. Sau đó, ta xây dựng hàm sau:

$$f(x) = \min_{1 \leq k \leq K} \left\{ \max \left\{ |c_k^T x + d_k|, \max_{1 \leq i \leq I_k} \{a_{ki}^T x + b_{ki}\} \right\} \right\}. \tag{2.2.1}$$

Vì hàm $\max$ và hàm giá trị tuyệt đối đều liên tục và tuyến tính từng phần, ta có thể xác nhận rằng hàm (2.2.1) là một hàm PWL liên tục. Tiếp theo, ta cần chứng minh rằng $\mathcal{B} = \{x : f(x) = 0\}$.

Đầu tiên, ta phải chứng minh $\mathcal{B} \subseteq \{x : f(x) = 0\}$. Lấy một điểm tùy ý trong $\mathcal{B}$, ký hiệu là x_0. Theo định nghĩa của tập hợp tuyến tính từng phần, ta có thể tìm một chỉ số k_0 sao cho $x_0 \in \Omega_{k_0}$ và $c_{k_0}^T x_0 + d_{k_0} = 0$. Lại có

$$\max_{1 \leq i \leq I_{k_0}} \{a_{k_0 i}^T x_0 + b_{k_0 i}\} \leq 0,$$

Do đó

$$\max \left\{ |c_k^T x + d_k|, \max_{1 \leq i \leq I_{k_0}} \{a_{ki}^T x + b_{ki}\} \right\} = 0,$$

trong đó I_{k_0} là số các ràng buộc xác định đa diện Ω_{k_0}. Vì $\Omega_1, \Omega_2, \ldots, \Omega_K$ tạo thành một phân hoạch của miền nên $\overset{\circ}{\Omega}_k \cap \overset{\circ}{\Omega}_l = \emptyset$, $\forall k \neq l$. Do đó, $x_0 \notin \overset{\circ}{\Omega}_k$,

$\forall k \neq k_0$, tức là

$$\max_{1\leq i\leq I_k}\{a_{ki}^T x_0 + b_{ki}\} \geq 0, \forall k \neq k_0,$$

từ đó suy ra

$$\begin{aligned}&\max\left\{|c_k^T x_0 + d_k|, \max_{1\leq i\leq I_k}\{a_{ki}^T x_0 + b_{ki}\}\right\}\\ &\quad\geq \max\left\{|c_{k_0}^T x_0 + d_{k_0}|, \max_{1\leq i\leq I_{k_0}}\{a_{k_0 i}^T x_0 + b_{k_0 i}\}\right\} = 0,\end{aligned}$$

và

$$f(x_0) = \max\left\{|c_{k_0}^T x_0 + d_{k_0}|, \max_{1\leq i\leq I_{k_0}}\{a_{k_0 i}^T x_0 + b_{k_0 i}\}\right\} = 0$$

Suy ra $\mathcal{B} \subseteq \{x : f(x) = 0\}$.

Tiếp theo, ta chứng minh rằng $\{x : f(x) = 0\} \subseteq B$. Giả sử $x_0 \in \{x : f(x) = 0\}$, tức là $f(x_0) = 0$. Khi đó tồn tại ít nhất một chỉ số $k_0 \in \{1, 2, \ldots, K\}$ sao cho

$$\max\left\{|c_{k_0}^T x_0 + d_{k_0}|, \max_{1\leq i\leq I_{k_0}}\{a_{k_0 i}^T x_0 + b_{k_0 i}\}\right\} = 0.$$

Rõ ràng, $c_{k_0}^T x_0 + d_{k_0} = 0$ và $a_{k_0 i}^T x_0 + b_{k_0 i} \leq 0, \forall 1 \leq i \leq I_{k_0}$, tức là $x_0 \in \Omega_{k_0}$. Từ đó, có thể kết luận rằng $x_0 \in \mathcal{B} \cap \Omega_{k_0} \subseteq \mathcal{B}$ và do đó $\{x : f(x) = 0\} \subseteq \mathcal{B}$.

Cuối cùng, ta được $\mathcal{B} = \{x : f(x) = 0\}$, tức là bất kỳ tập hợp tuyến tính từng phần nào cũng có thể được biểu diễn như là tập nghiệm của một phương trình tuyến tính từng phần liên tục. □

2.2.2. Phương trình tuyến tính từng phần

Từ các công thức $|t| = \max\{t, -t\}$, $\forall t \in \mathbb{R}$ và $\max\{t_1, \max\{t_2, t_3\}\} = \max\{t_1, t_2, t_3\}$, $\forall t_1, t_2, t_3 \in \mathbb{R}$, ta viết lại (2.2.1) dưới dạng công thức min-

max như sau:

$$\min_{1\leq k\leq K}\left\{\max\{c_k^T x+d_k,-c_k^T x-d_k,a_{k1}^T x+b_{k1},\ldots,a_{kI_k}^T x+b_{kI_k}\}\right\}.$$

Sử dụng công thức min-max, một bộ phân loại tuyến tính từng phần có thể được xây dựng. Vấn đề xác định các tham số có thể được đặt ra như một bài toán tối ưu không lồi và không khả vi của việc tối thiểu hóa hàm mất mát của các điểm bị phân loại sai. Tuy nhiên, vì bài toán tối ưu liên quan rất khó giải, số lượng các vùng con bị giới hạn ở một số nhỏ. Ví dụ, trong [8], chỉ các trường hợp với $K \leq 5$ được xem xét trong thí nghiệm số. Để thu được các tham số một cách hiệu quả và đạt được khả năng tổng quát hóa tốt, trong đề án này, chúng tôi áp dụng kỹ thuật **Máy vector hỗ trợ (SVM)**. Để xây dựng một SVM mong muốn, ta cần một công thức khác được biến đổi từ (2.2.1) dựa trên những điều sau:

Bổ đề 2.2.1 (Định lý 1, Wang and Sun (18))**.** *Cho hàm số* $f(x):\mathbb{R}^n\to\mathbb{R}$

$$f(x)=\max_{1\leq k\leq K}\left\{\min_{1\leq i\leq I_k}\{a_{ki}^T x+b_{ki}\}\right\}$$

tồn tại M *hàm cơ sở* $\phi_m(x)$ *với các tham số* $w_m\in\mathbb{R}$, $p_{mi}\in\mathbb{R}^n$ *và* $q_{mi}\in\mathbb{R}$ *sao cho*

$$f(x)=\sum_{m=1}^{M} w_m\phi_m(x),$$

trong đó

$$\phi_m(x)=\max\{p_{m0}^T x+q_{m0},p_{m1}^T x+q_{m1},\ldots,p_{mn}^T x+q_{mn}\}.$$

Theo Bổ đề 2.2.1, Định lý 2.2.1 và công thức $\min_k \max_i\{t_{ik}\} = -\max_k \min_i\{-t_{ik}\}$, ta nhận được một công thức các hàm số tuyến tính từng phần khác. Kết quả này được mô tả trong Định lý sau đây, điều này làm cho SVM có thể áp dụng để xây dựng các bộ phân loại tuyến tính từng phần.

Định lý 2.2.2. *Mọi tập tuyến tính từng phần $\mathcal{B}$ đều có thể được biểu diễn như là nghiệm của phương trình tuyến tính từng phần, tức là $\mathcal{B} = \{x : f(x) = 0\}$, trong đó, $f(x)$ thỏa*

$$f(x) = \sum_{m=1}^{M} w_m \phi_m(x), \tag{2.2.2}$$

và

$$\phi_m(x) = \max\{p_{m0}^T x + q_{m0}, p_{m1}^T x + q_{m1}, \ldots, p_{mn}^T x + q_{mn}\}. \tag{2.2.3}$$

2.2.3. SVM với ánh xạ đặc trưng tuyến tính từng phần

Biểu diễn bộ phân loại tuyến tính từng phần như một tổ hợp tuyến tính của các hàm cơ sở cho phép sử dụng kỹ thuật SVM để xác định các hệ số tuyến tính của (2.2.2). SVM với ánh xạ đặc trưng tuyến tính từng phần cũng có thể được coi là một mạng nơ-ron nhiều lớp (MLP) với các lớp ẩn. Mối quan hệ giữa ánh xạ đặc trưng của SVM và lớp ẩn của MLP đã được mô tả trong [20]. Sử dụng hàm tuyến tính từng phần (2.2.3) làm ánh xạ đặc trưng, chúng ta có thể thiết lập một SVM cung cấp biên phân loại tuyến tính từng phần. Ký hiệu $\tilde{p}_{mi} = p_{mi} - p_{m0}$, $\tilde{q}_{mi} = q_{mi} - q_{m0}$. Công thức (2.2.3) có thể được biến đổi

tương đương thành

$$\phi_m(x) = p_{m0}^T x + q_{m0} + \max\{0, \tilde{p}_{m1}x + \tilde{q}_{m1}, \ldots, \tilde{p}_{mn}x + \tilde{q}_{mn}\}.$$

Ký hiệu thành phần thứ i của x là $x(i)$, chúng ta sử dụng ánh xạ đặc trưng tuyến tính từng phần sau đây trong không gian n chiều:

$$\phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_M(x)]^T,$$

trong đó

$$\phi_m(x) = \begin{cases} x(m), & m = 1, \ldots, n, \\ \max\{0, p_{m1}^T x + q_{m1}, \ldots, p_{mn}^T x + q_{mn}\}, & m = n+1, \ldots, M. \end{cases} \tag{2.2.4}$$

Ta có thể xây dựng một loạt các SVM với ánh xạ đặc trưng tuyến tính từng phần, được gọi là PWL-SVMs. Ví dụ, một ánh xạ đặc trưng tuyến tính từng phần có thể được áp dụng trong C-SVM [21] và ta có công thức sau, được gọi là PWL-C-SVM:

$$\min_{\mathrm{w},\mathrm{w}_0,e} \quad \frac{1}{2}\sum_{m=1}^{M} \mathrm{w}_m^2 + \gamma \sum_{k=1}^{N} e_k$$

sao cho:

$$y_k \left[\mathrm{w}_0 + \sum_{m=1}^{M} \mathrm{w}_m \phi_m(x_k) \right] \geq 1 - e_k, \quad \forall k,$$

$$e_k \geq 0, \quad k = 1, 2, \ldots, N, \tag{2.2.5}$$

trong đó $x_k \in \mathbb{R}^n, k = 1, 2, \ldots, N$ là các dữ liệu đầu vào, $y_k \in \{+1, -1\}$ là

các nhãn tương ứng, và ánh xạ đặc trưng $\phi(x) = [\phi_1(x), \phi_2(x), \ldots, \phi_M(x)]^T$ có dạng như trong (2.2.4). Để tránh nhầm lẫn với các tham số trong ánh xạ đặc trưng, trong đề án này, ta sử dụng ký hiệu w_0 để chỉ *bias term* trong SVMs. Bài toán đối ngẫu là:

$$\max_{\alpha} \quad -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N} y_k y_l \kappa(x_k, x_l)\alpha_k\alpha_l + \sum_{k=1}^{N}\alpha_k$$

sao cho

$$\sum_{k=1}^{N} \alpha_k y_k = 0, \quad 0 \leq \alpha_k \leq \gamma, \quad k = 1, 2, \ldots, N. \tag{2.2.6}$$

trong đó α là biến đối ngẫu và hạt nhân là

$$\kappa(x_k, x_l) = \phi(x_k)^T\phi(x_l) = \sum_{m=1}^{M}\phi_m(x_k)\phi_m(x_l). \tag{2.2.7}$$

Từ bài toán nguyên thủy (2.2.5), ta có bộ phân loại tuyến tính từng phần

$$\text{sign}\left(\mathrm{w}^T\phi(x) + \mathrm{w}_0\right). \tag{2.2.8}$$

Số lượng biến trong (2.2.5) là $M + N + 1$, trong khi ở bài toán đối ngẫu (2.2.6) số lượng biến là N, do đó ta ưu tiên giải (2.2.6) để có bộ phân loại sau

$$\text{sign}\left(\sum_{k=1}^{N}\alpha_k y_k \kappa(x, x_k) + \mathrm{w}_0\right).$$

Để xây dựng bộ phân loại sử dụng dạng đối ngẫu trên, cần lưu trữ $\alpha_k, x_k, \mathrm{w}_0$ và đối với (2.2.8) ta chỉ cần nhớ $\mathrm{w}_m, \mathrm{w}_0$. Do đó, ta giải (2.2.6) để thu được

các biến đối ngẫu và sau đó tính w bằng

$$\mathrm{w}_m = \sum_{k=1}^{N} \alpha_k y_k \phi_m(x_k)$$

và sử dụng (2.2.8) để phân loại.

Sử dụng ánh xạ đặc trưng tuyến tính từng phần trong SVM, ta có một bộ phân loại cung cấp biên phân loại tuyến tính từng phần và hưởng lợi từ các ưu điểm của SVMs. Trong [13], các nhà nghiên cứu đã xây dựng một bộ phân loại tuyến tính từng phần sử dụng SVM và thu được kết quả tốt. Tuy nhiên, phương pháp của họ chỉ xử lý được các trường hợp phân tách được và còn một số vấn đề quan trọng chưa được giải quyết, bao gồm *"làm thế nào để định nghĩa biên mềm (soft margins) hợp lý?"* và *"làm thế nào để mở rộng cho tập dữ liệu không phân tách?"* như đã đề cập trong [13]. Sử dụng ánh xạ đặc trưng tuyến tính từng phần, chúng ta đã thành công xây dựng một PWL-SVM, cung cấp biên phân loại tuyến tính từng phần và có thể xử lý bất kỳ loại dữ liệu nào.

Tương tự, ta có thể sử dụng ánh xạ đặc trưng tuyến tính từng phần trong SVM bình phương tối thiểu (LS-SVM [22–24]) và thu được PWL-LS-SVM sau:

$$\min_{\mathrm{w},\mathrm{w}_0,e} \quad \frac{1}{2}\sum_{m=1}^{M} \mathrm{w}_m^2 + \gamma \frac{1}{2}\sum_{k=1}^{N} e_k^2$$

sao cho:

$$y_k \left[\mathrm{w}_0 + \sum_{m=1}^{M} \mathrm{w}_m \phi_m(x_k) \right] = 1 - e_k, \quad k = 1, 2, \ldots, N. \tag{2.2.9}$$

Bài toán đối ngẫu của (2.2.9) là một phương trình tuyến tính của α, w_0, tức là

$$\begin{bmatrix} 0 & y^T \\ y & K + \frac{\mathbf{I}}{\gamma} \end{bmatrix} \begin{bmatrix} \mathrm{w}_0 \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{1} \end{bmatrix}, \tag{2.2.10}$$

trong đó $\mathbf{I} \in \mathbb{R}^{N \times N}$ là ma trận đơn vị, $\mathbf{1} \in \mathbb{R}^N$ là vector với các thành phần bằng một, $\alpha \in \mathbb{R}^N$ là biến đối ngẫu, $K_{kl} = y_k y_l \kappa(x_k, x_l)$ và hạt nhân giống như (2.2.7). Số lượng biến liên quan trong bài toán nguyên thủy (2.2.9) và bài toán đối ngẫu (2.2.10) là $M + 1$ và $N + 1$, tương ứng. Do đó, chúng ta ưu tiên giải (2.2.9) để xây dựng bộ phân loại khi $M \leq N$. Ngược lại, tức là khi $M > N$, chúng ta giải (2.2.10) để thu được các biến đối ngẫu α_k, sau đó tính các hệ số w_m và sử dụng dạng nguyên thủy làm bộ phân loại.

2.2.4. Khả năng phân loại của ánh xạ đặc trưng tuyến tính từng phần

SVM với ánh xạ đặc trưng tuyến tính từng phần (PWL) tạo ra một ranh giới tuyến tính từng phần và thừa hưởng các đặc tính tốt của SVM. Khả năng phân loại của PWL-SVM liên quan đến công thức cụ thể của $\phi_m(x)$.
Công thức đơn giản nhất là

$$\phi_m(x) = \max\{0, x(i_m) - q_m\}, \tag{2.2.11}$$

trong đó $q_m \in \mathbb{R}$ và $i_m \in \{1, \ldots, n\}$ biểu thị thành phần được sử dụng trong $\phi_m(x)$. Sử dụng (2.2.11) làm ánh xạ đặc trưng, một bộ phân loại tuyến tính từng phần cộng có thể được xây dựng. Các điểm thay đổi của ranh giới phân loại tuyến tính từng phần nên được đặt tại các ranh giới của các vùng con và các ranh giới được cung cấp bởi (2.2.11) bị giới hạn là các đường song song với một trong các trục.

Vì các ranh giới của các vùng con được định nghĩa bởi (2.2.11) không đủ linh hoạt, một số bộ phân loại mong muốn không thể đạt được. Để mở rộng khả năng phân loại, ta nên mở rộng (2.2.11) thành

$$\phi_m(x) = \max\{0, p_m^T x - q_m\}, \tag{2.2.12}$$

trong đó $p_m \in \mathbb{R}^n, q_m \in \mathbb{R}$. Công thức này được gọi là một siêu phẳng bản lề - Hinging Hyper Plane(HH [16]) và các ranh giới của các vùng con được cung cấp bởi siêu phẳng là các đường xuyên suốt miền, linh hoạt hơn so với (2.2.11). Để có được một bộ phân loại tuyến tính từng phần với khả năng phân loại mạnh mẽ hơn, ta có thể thêm nhiều hàm tuyến tính theo cách sau

$$\phi_m(x) = \max\{0, p_{m1}^T x - q_{m1}, p_{m2}^T x - q_{m2}, \ldots\}.$$

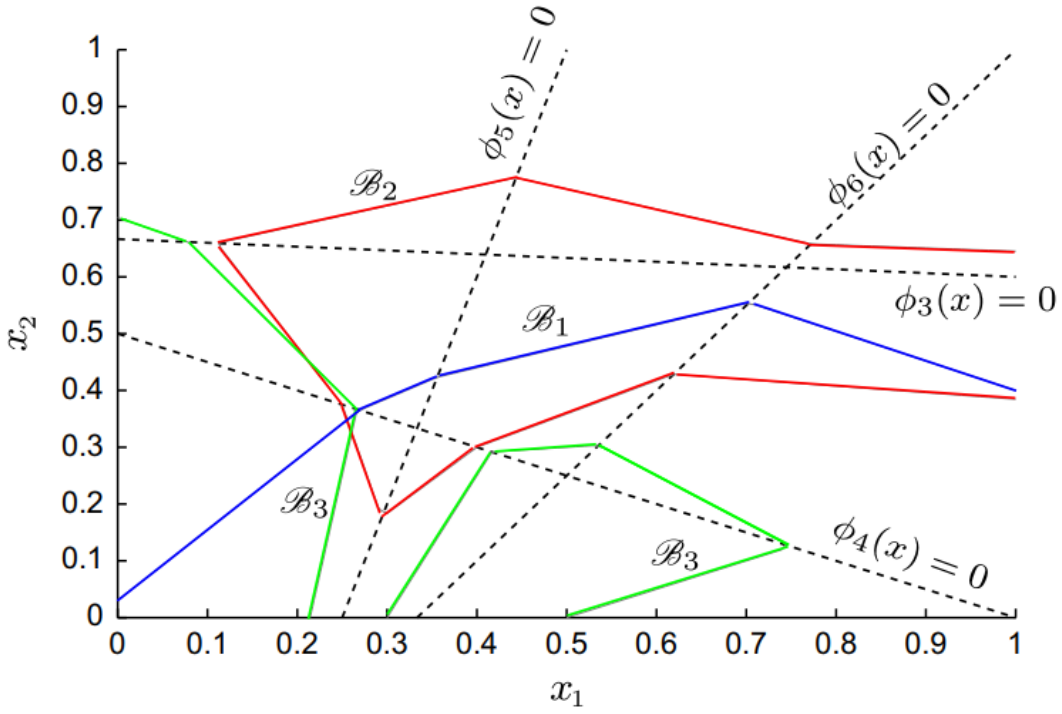
Khả năng phân loại được mở rộng cùng với sự gia tăng của các hàm tuyến tính được sử dụng trong $\max\{\}$. Như đã chứng minh trong Định lý 2.2.2, một ranh giới tuyến tính từng phần bất kỳ trong không gian n-chiều có thể được thực hiện bằng cách sử dụng n hàm tuyến tính, tức là

$$\phi_m(x) = \max\{0, p_{m1}^T x - q_{m1}, p_{m2}^T x - q_{m2}, \ldots, p_{mn}^T x - q_{mn}\}, \tag{2.2.13}$$

Theo ký hiệu trong [18], chúng ta gọi (2.2.13) là ánh xạ đặc trưng siêu phẳng bản lề tổng quát - Generalized Hinging HyperPlane (GHH).

2.2.5. Tham số trong ánh xạ đặc trưng tuyến tính từng phần

Giống như các ánh xạ đặc trưng phi tuyến hoặc các hạt nhân khác, ánh xạ đặc trưng tuyến tính từng phần có một số tham số phi tuyến, có ảnh hưởng



Hình 3: Các ranh giới của các vùng con cho (2.2.11) và các ranh giới phân loại liên quan.

lớn đến hiệu suất phân loại nhưng khó điều chỉnh tối ưu. Để có được các tham số hợp lý cho ánh xạ đặc trưng tuyến tính từng phần, ta nghiên cứu ý nghĩa hình học của các tham số. Từ định nghĩa của một tập tuyến tính từng phần, miền được chia thành các vùng con Ω_k, trong mỗi vùng con đó tập tuyến tính từng phần bằng một siêu phẳng. Nói chung, các tham số trong ánh xạ đặc trưng tuyến tính từng phần xác định cấu trúc vùng con và siêu phẳng trong mỗi vùng con được thu được bằng kỹ thuật SVM.

Xét lại tập dữ liệu hai mặt trăng được hiển thị trong Hình 1. Ranh giới bao gồm ba đoạn nằm trong vùng con Ω_k như minh họa trong Hình 2. Để xây dựng bộ phân loại mong muốn, ta đặt

$$\phi_1(x) = x(1), \quad \phi_2(x) = x(2),$$

$$\phi_3(x) = \max\{0, x(2) - \frac{1}{3}\}, \quad \phi_4(x) = \max\{0, x(2) - \frac{2}{3}\},$$

tức là, ta sử dụng ánh xạ đặc trưng (2.2.11) với $i_1 = 1, q_1 = 0, i_2 = 2, q_2 = 0, i_3 = 2, q_3 = \frac{1}{3}$, và $i_4 = 2, q_4 = \frac{2}{3}$. Tiếp theo giải PWL-C-SVM hoặc PWL-LS-SVM dẫn đến $\mathrm{w}_0, \ldots, \mathrm{w}_4$, xác định một ranh giới tuyến tính từng phần $\mathrm{w}_0 + \sum_{m=1}^{4} \mathrm{w}_m f_m(x) = 0$. Trong ví dụ này, dựa vào một số kiến thức trước đó đã được biết, các tham số hợp lý cho một ánh xạ đặc trưng tuyến tính từng phần có thể được tìm thấy một cách hiệu quả. Trong các vấn đề hộp đen, chúng ta đặt các tham số của (2.2.11) bằng cách chia đều miền trên mỗi trục thành nhiều đoạn.

Đối với các ánh xạ đặc trưng tuyến tính từng phần khác, chúng ta sử dụng các tham số ngẫu nhiên. Các ranh giới của các đoạn được cung cấp bởi (2.2.12) là các siêu phẳng $p_m^T x + q_m = 0$. Để có được $p_m \in \mathbb{R}^n$ và $q_m \in \mathbb{R}$, trước tiên chúng ta tạo ra n điểm trong miền với phân phối đều, sau đó chúng ta chọn $p_m(1)$ từ $\{1, -1\}$ với xác suất bằng nhau, và tính q_m và các thành phần khác của p_m sao cho các điểm được tạo ra nằm trong $p_m^T x + q_m = 0$. Các tham số thu được theo cách này có thể cung cấp các ranh giới phân loại linh hoạt.

Hình 3 cho thấy các ranh giới của các vùng con cho liên quan đến ánh xạ đặc trưng tuyến tính tưng phần được tạo ngẫu nhiên. Bốn đường nét đứt tương ứng với $f_3(x) = 0$, $f_4(x) = 0$, $f_5(x) = 0$, và $f_6(x) = 0$, tương ứng. Ranh giới phân loại là $\mathcal{B} = \{x : w_0 + \sum_{m=1}^{6} w_m f_m(x) = 0\}$. $\mathcal{B}_1$ với $w_0 = -1.5, w_1 = 0.5, w_2 = 0.9, w_3 = 0.7, w_4 = 0.8, w_5 = 0.1, w_6 = 0.33$, $\mathcal{B}_2$ với $w_0 = -0.5, w_1 = -0.5, w_2 = 1, w_3 = -1, w_4 = 0.8, w_5 = -0.25, w_6 = 0.33$, và $\mathcal{B}_3$ với $w_0 = -0.5, w_1 = 1, w_2 = -2, w_3 = 1, w_4 = 4, w_5 = -0.75, w_6 = 0.67$ được minh họa bằng các đường màu đỏ, xanh dương, và xanh lá cây, cho thấy sự linh hoạt của ranh giới phân loại, có thể là lồi ($\mathcal{B}_1$), không lồi ($\mathcal{B}_2$), và

không kết nối ($\mathcal{B}_3$).

$$
\begin{aligned}
&f_1(x) = x(1), \quad f_2(x) = x(2),\\
&f_3(x) = \max\{0, x(1) + 1.5x(2) - 1.0\},\\
&f_4(x) = \max\{0, x(1) + 2x(2) - 1.0\},\\
&f_5(x) = \max\{0, -x(1) + 0.25x(2) + 0.25\},\\
&f_6(x) = \max\{0, -x(1) + 0.67x(2) + 0.33\}.
\end{aligned} \tag{2.2.14}
$$

Ba ranh giới phân loại có thể tương ứng với các nhóm w_m khác nhau được hiển thị. Có thể thấy sự linh hoạt của các bộ phân loại tiềm năng, từ đó bộ phân loại tối ưu có thể được chọn ra bằng SVM. Trong hầu hết các trường hợp, hiệu suất phân loại là thỏa đáng, nếu không, chúng ta có thể tạo ra một nhóm tham số khác. Tương tự, các tham số của (2.2.13) có thể được tạo ra và các vùng con kết quả cung cấp các ranh giới phân loại linh hoạt.