**Thuật toán SVM**

1. **Tại sao nên sử dụng svm để phân lớp dữ liệu**

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

1. **Thuật toán SVM**
2. Định nghĩa

**SVM** là một thuật toán giám sát, nó có thể sử dụng cho cả việc phân loại hoặc đệ quy. Tuy nhiên nó được sử dụng chủ yếu cho việc phân loại.

Là phương pháp dựa trên nền tảng của lý thuyết thống kê nên có một nền tảng toán học chặt chẽ để đảm bảo rằng kết quả tìm được là tối ưu

- Là 1 phương pháp thử nghiệm,đưa ra 1 trong những phương pháp mạnh và chính xác nhất trong số các thuật toán nổi tiếng về phân lớp dữ liệu

1. Phân loại

* Linear SVM

Linear SVM được sử dụng cho bộ dữ liệu có thể phân tách tuyến tính, có nghĩa là nếu một tập dữ liệu có thể được phân loại thành 2 lớp bằng cách sử dụng một đường thẳng(hoặc siêu phẳng) duy nhất, thì dữ liệu đó được gọi là dữ liệu có thể phân tách tuyến tính. Và bộ phân loại sử dụng được là bộ phân loại SVM tuyến tính. 

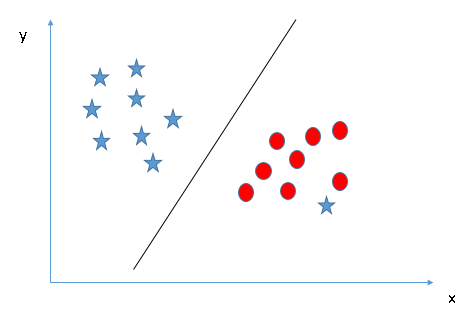
* Non-Linear SVM

Non-Linear SVM được sử dụng cho dữ liệu được phân tách phi tuyến tính, có nghĩa là nếu một tập dữ liệu không thể được phân loại bằng cách sử dụng một đường thẳng, thì dữ liệu đó được gọi là dữ liệu phi tuyến tính. Và bộ phân loại sử dụng được gọi là bộ phân loại SVM phi tuyến tính.

D

1. Cơ sở lí thuyết

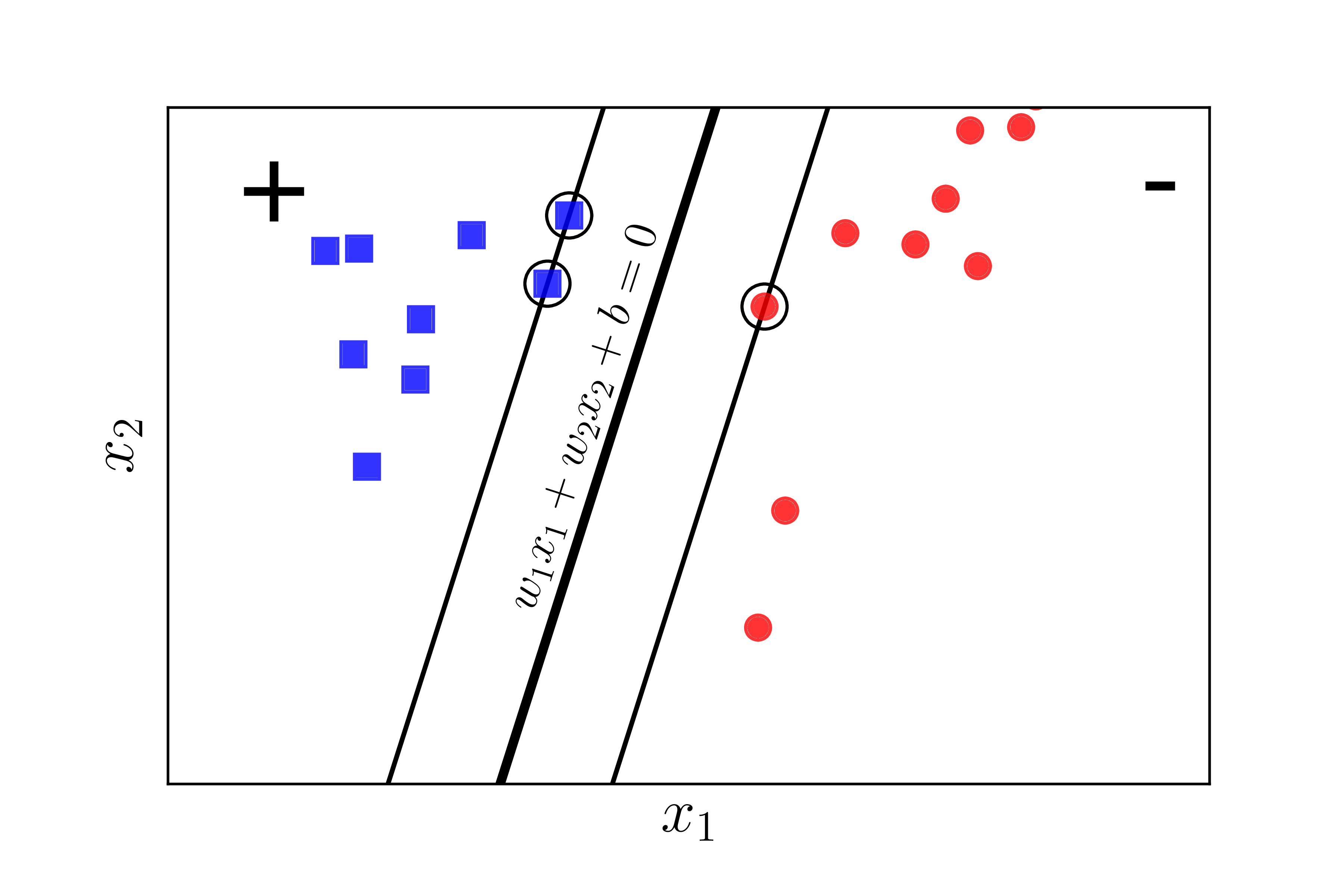
Trong thuật toán này, chúng ta vẽ đồi thị dữ liệu là các điểm trong n chiều ( ở đây n là số lượng các tính năng bạn có) với giá trị của mỗi tính năng sẽ là một phần liên kết. Sau đó chúng ta thực hiện tìm "đường bay" (hyper-plane) phân chia các lớp. Hyper-plane nó chỉ hiểu đơn giản là 1 đường thẳng có thể ph**â**n chia các lớp ra thành hai phần riêng biệt.



1. Xây dựng bài toán tối ưu cho SVM:

Giả sử rằng các cặp dữ liệu của training set là (, (, … ( với vector thể hiện đầu vào của một điểm dữ liệu và là nhãn của điểm dữ liệu đó. d là số chiều của dữ liệu và n là số điểm dữ liệu. Giả sử rằng nhãn của mỗi điểm dữ liệu được xác định bởi =1 (class 1) hoặc (class 2).

Để giúp các bạn dễ hình dung, chúng ta cùng xét trường hợp trong không gian hai chiều dưới đây. Không gian hai chiều để các bạn dễ hình dung, các phép toán hoàn toàn có thể được tổng quát lên không gian nhiều chiều.

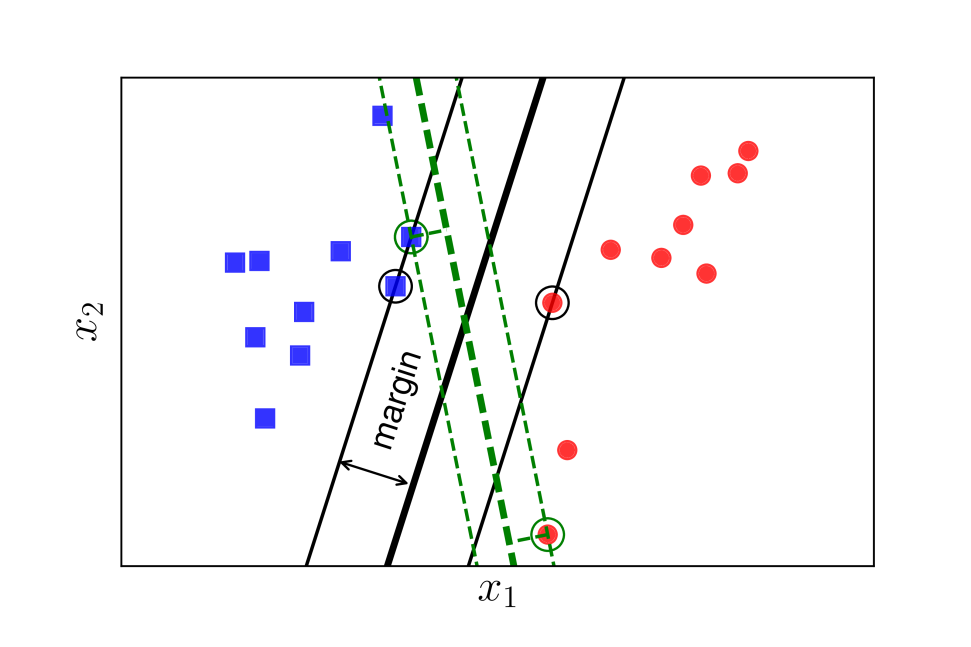


Giả sử rằng các điểm vuông xanh thuộc class 1, các điểm tròn đỏ thuộc class -1 và mặt ( là mặt phân chia giữa hai classes. Hơn nữa, class 1 nằm về *phía dương*, class -1 nằm về *phía âm* của mặt phân chia. Nếu ngược lại, ta chỉ cần đổi dấu của w và b. Chú ý rằng chúng ta cần đi tìm các hệ số w và b

Ta quan sát thấy một điểm quan trọng sau đây: với cặp dữ liệu ( , ) bất kỳ, khoảng cách từ điểm đó tới mặt phân chia là:

Điều này có thể dễ nhận thấy vì theo giả sử ở trên,  luôn cùng dấu với *phía* của . Từ đó suy ra  cùng dấu với (), và tử số luôn là 1 số không âm.

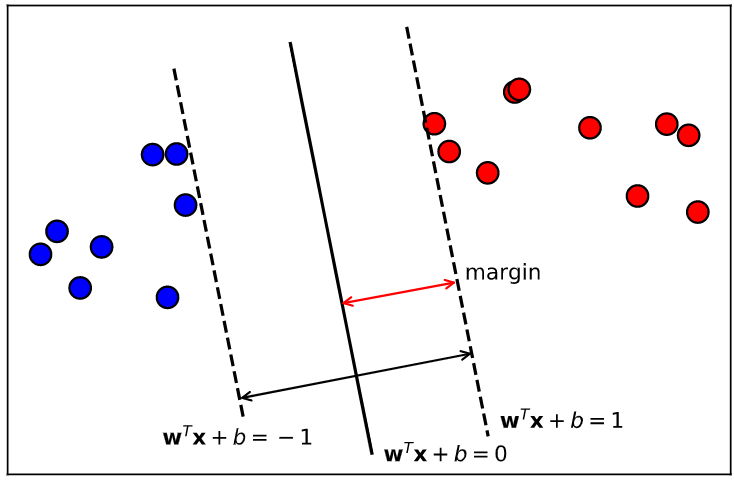
Với mặt phần chia như trên, margin được tính là khoảng cách gần nhất từ 1 điểm tới mặt đó (bất kể điểm nào trong hai classes):



margin = min

Bài toán tối ưu trong SVM chính là bài toán tìm w và b sao cho *margin* này đạt giá trị lớn nhất:

(W, b) = =



Sau khi tìm được mặt phân cách : . Class của bất kỳ điểm nào sẽ được xác định bằng cách:

Class(x) =sgn(

Trong đó hàm sgn là xác định dấu, nhận giá trị 1 nếu đối số là không âm và -1 nếu ngược lại

1. Ưu nhược điểm của thuật toán SVM

Ưu điểm:

Là một kĩ thuật phân lớp khá phổ biến, SVM thể hiện được nhiều ưu điểm trong số đó có việc tính toán hiệu quả trên các tập dữ liệu lớn.

Xử lý trên không gian số chiều cao: SVM là một công cụ tính toán hiệu quả trong không gian chiều cao, trong đó đặc biệt áp dụng cho các bài toán phân loại văn bản và phân tích quan điểm nơi chiều có thể cực kỳ lớn.

Tiết kiệm bộ nhớ: Do chỉ có một tập hợp con của các điểm được sử dụng trong quá trình huấn luyện và ra quyết định thực tế cho các điểm dữ liệu mới nên chỉ có những điểm cần thiết mới được lưu trữ trong bộ nhớ khi ra quyết định.

Tính linh hoạt - phân lớp thường là phi tuyến tính. Khả năng áp dụng Kernel mới cho phép linh động giữa các phương pháp tuyến tính và phi tuyến tính từ đó khiến cho hiệu suất phân loại lớn hơn.

Nhược điểm:

* Bài toán số chiều cao: Trong trường hợp số lượng thuộc tính (p) của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu (n) thì SVM cho kết quả khá tồi.
* Chưa thể hiện rõ tính xác suất: Việc phân lớp của SVM chỉ là việc cố gắng tách các đối tượng vào hai lớp được phân tách bởi siêu phẳng SVM. Điều này chưa giải thích được xác suất xuất hiện của một thành viên trong một nhóm là như thế nào. Tuy nhiên hiệu quả của việc phân lớp có thể được xác định dựa vào khái niệm margin từ điểm dữ liệu mới đến siêu phẳng phân lớp mà chúng ta đã bàn luận ở trên.

1. **Ứng dụng**

Như chúng ta đã thấy, SVM phụ thuộc vào thuật toán học có giám sát. Mục đích của việc sử dụng SVM là phân loại chính xác dữ liệu chưa nhìn thấy, SVM đã được ứng dụng trong nhiều lĩnh vực. Một số ứng dụng phổ biến của SVM là:

Face detection(nhận diện khuôn mặt): SVMs phân loại các phần của hình ảnh thành khuôn mặt và không phải khuôn mặt và tạo đường bao hình vuông xung quanh khuôn mặt.

Text and hypertext categorization(Phân loại văn bản và siêu văn bản): cho phép phân loại văn bản và siêu văn bản cho cả mô hình quy nạp và mô hình chuyển đổi. Họ sử dụng dữ liệu training để phân loại tài liệu thành các loại khác nhau. Nó phân loại dựa trên điểm số được tạo ra và sau đó so sánh với giá trị ngưỡng.

Classification of images(phân loại hình ảnh): Sử dụng SVMs cung cấp độ chính xác tìm kiếm tốt hơn để phân loại hình ảnh. Nó cung cấp độ chính xác cao hơn so với các kỹ thuật tìm kiếm dựa trên truy vấn truyền thống.

Handwriting recognition: sử dụng SVM để nhận dạng các ký tự viết tay được sử dụng rộng rãi.