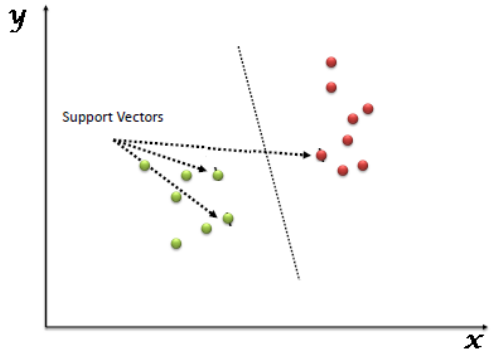
GIẢI THUẬT SUPPORT VECTOR MACHINE

1. Support Vector Machine là gì?

Support Vector Machine (SVM) là một giải thuật máy học có giám sát được sử dụng phổ biến cho các bài toán phân lớp. Trong giải thuật, mỗi hạng mục dữ liệu (data item) được biểu diễn dưới dạng một điểm trong không gian n chiều (n thường là số đặc trưng và giá trị của mỗi đặc trưng là giá trị của một tọa độ cụ thể). Sau đó, dữ liệu được phân lớp bằng cách tìm ra một siêu phẳng (hyper-plane) để tách biệt 2 lớp là tốt nhất.

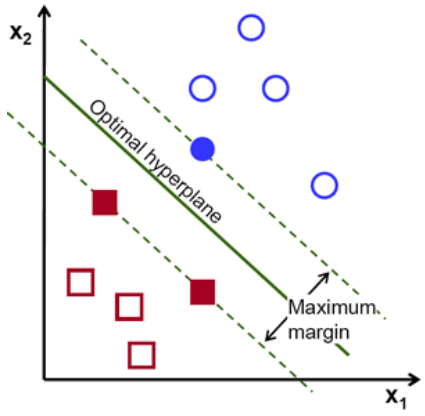


Hình 1: Dữ liệu được phân lớp với SVM

2. Nội dung của giải thuật

Cho trước một tập dữ liệu huấn luyện được biểu diễn trong không gian vector, trong đó mỗi hạng mục dữ liệu được xem là một điểm trong không gian này. Giải thuật này là xây dựng một siêu phẳng tốt nhất có thể phân chia các điểm trong không gian này thành hai lớp riêng biệt tương ứng. Chất lượng của siêu phẳng này được quyết định bởi một khoảng cách (được gọi là biên – margin) từ điểm dữ liệu gần nhất đến siêu phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt, đồng thời việc phân loại càng chính xác. Nhờ vậy, SVM có thể giảm thiểu việc phân lớp sai (misclassification) đối với điểm dữ liệu mới đưa vào.

Mục đích của giải thuật SVM là tìm được khoảng cách biên lớn nhất để cho ra kết quả phân lớp tốt nhất.



Hình 2: Siêu phẳng phân chia dữ liệu huấn luyện thành 2 lớp với khoảng cách biên lớn nhất. Các điểm gần nhất (điểm tô màu) là các Support Vector

3. Siêu phẳng tốt nhất được tính như thế nào?

Một siêu phẳng sẽ có phương trình: $f(x) = \beta_0 + \beta^T x$, trong đó β là vector trọng số (weight vector) và β_0 là bias.

Bằng việc thay đổi β và β_0 sẽ có vô số cách biểu diễn siêu phẳng tốt nhất, trong số đó có một lựa chọn để biểu diễn cho siêu phẳng là:

$$|\beta_0 + \beta^T x| = 1$$

Trong đó x đại diện cho các mẫu dữ liệu huấn luyện gần nhất với siêu phẳng. Tổng quát, các mẫu dữ liệu huấn luyện gần với siêu phẳng chính là các *support vector*. Sự biểu diễn này được gọi là *canonical hyperplane*.

Khi đó, khoảng giữa các điểm mẫu x và siêu phẳng (β, β_0) được xác định:

$$\text{distance} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|}.$$

Từ đó, suy ra được khoảng cách giữa các điểm support vector và siêu phẳng canonical hyperplane là:

$$\text{distance}_{\text{support vectors}} = \frac{|\beta_0 + \beta^T x|}{\|\beta\|} = \frac{1}{\|\beta\|}. \quad (*)$$

Nhắc lại, mục đích của SVM là tìm được khoảng cách biên (margin) lớn nhất. Do đó, từ (*) ta xác định được biên, ký hiệu M, bằng 2 lần khoảng cách đến các điểm mẫu gần nhất:

$$M = \frac{2}{\|\beta\|}$$

Như vậy, bài toán tìm biên M lớn nhất tương đương với bài toàn cực tiểu hàm $L(\beta)$ thỏa mãn một số ràng buộc như sau:

$$\min_{\beta,\beta_0} L(\beta) = \frac{1}{2}\|\beta\|^2 \text{ subject to } y_i(\beta^T x_i + \beta_0) \geq 1 \ \forall i,$$

Với x_i biểu diễn cho tất cả các mẫu dữ liệu huấn luyện, y_i biểu diễn cho mỗi nhãn (label) của các mẫu huấn luyện. Đây là bài toán tối ưu Lagrange có thể được giải bằng việc dùng nhân tử Lagrange (Lagrange multipliers) để tìm được β và β0 của siêu phẳng.

Xác định phân lớp cho một điểm dữ liệu mới: Sau khi tìm được siêu phẳng tốt nhất: $\beta_0 + \beta^T x$, phân lớp của bất kỳ một điểm nào sẽ được xác định đơn giản bằng cách:

$$\text{class}(x) = \text{sgn}(\ \beta_0 + \beta^T x\ )$$

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

4. Ưu và nhược điểm

Ưu điểm:

- Tính toán hiệu quả trên các tập dữ liệu lớn
- 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 dịnh
- 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:

- Trong trường hợp số lượng thuộc tính của tập dữ liệu lớn hơn rất nhiều so với số lượng dữ liệu 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.

5. Kết luận

SVM là một phương pháp hiệu quả cho bài toán phân lớp dữ liệu. Nó là một công cụ đắc lực cho các bài toán về xử lý ảnh, phân loại văn bản, phân tích quan điểm.