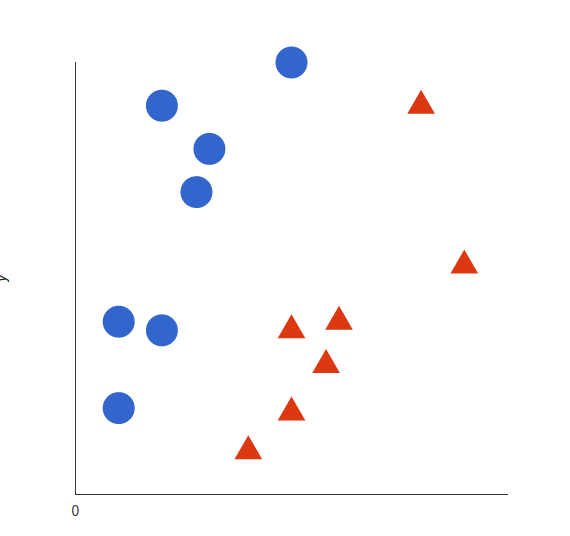
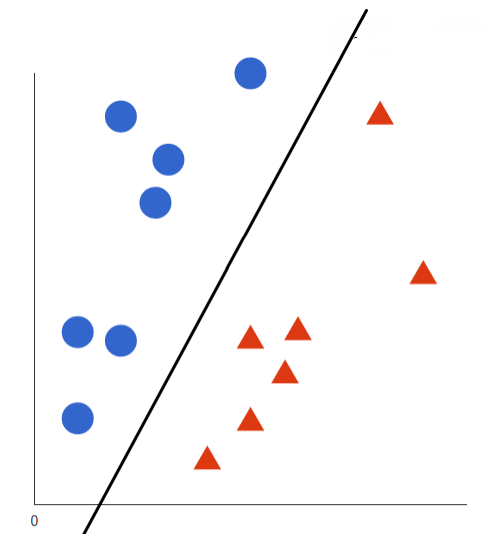
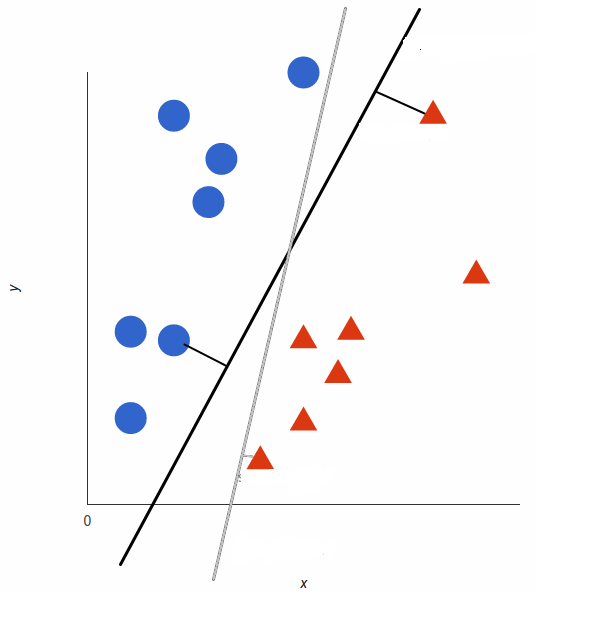
1. Hoạt động của SVM

Khái niệm cơ bản về SVM và cách hoạt động của nó có thể được khái quát trong ví dụ sau. Hãy tưởng tượng chúng ta có nhãn: đỏ và xanh cùng dữ liệu đầu vào gồm hai tọa độ: x và y. Ta sẽ muốn phân loại, với một cặp tọa độ (x, y), sẽ đưa ra kết luận nó có màu đỏ hay xanh. Chúng ta vẽ biểu đồ dữ liệu huấn luyện đã được gắn nhãn trên một mặt phẳng như sau:

SVM sẽ lấy các điểm dữ liệu này và đưa lên mặt phẳng và tìm một đường phân chia tối ưu nhất. Đường này là ranh giới quyết định: bất cứ điểm nào nằm về bên trái sẽ được nhận định có màu xanh và ngược lại.

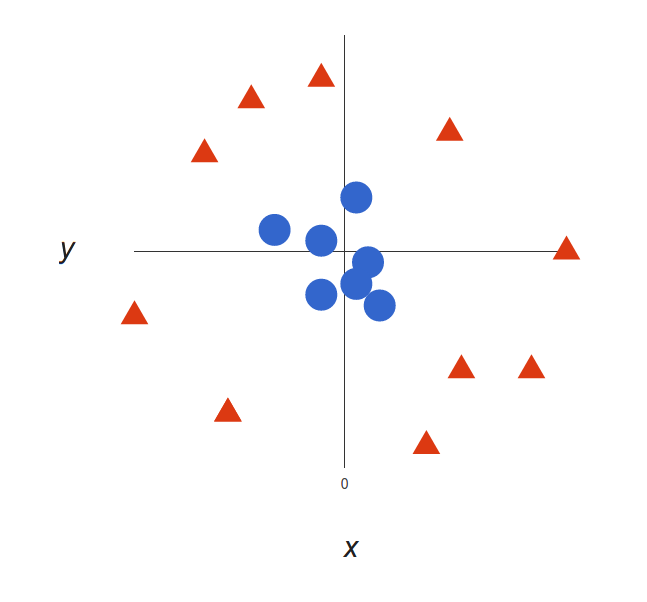


Vậy đường thẳng nào sẽ là tối ưu nhất? Đối với SVM, đường thẳng tối ưu sẽ là đường đảm bảo khoảng cách giữa điểm gần nhất với đường thẳng của mỗi phần sẽ là lớn nhất có thể

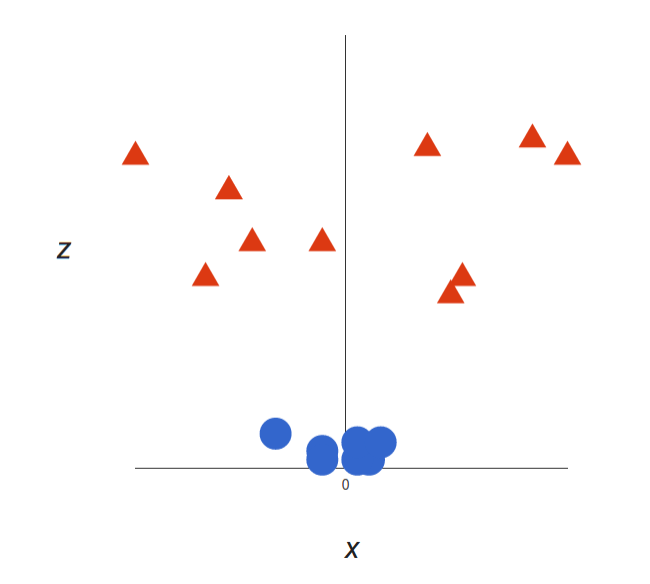


2. Dữ liệu phi tuyến tính

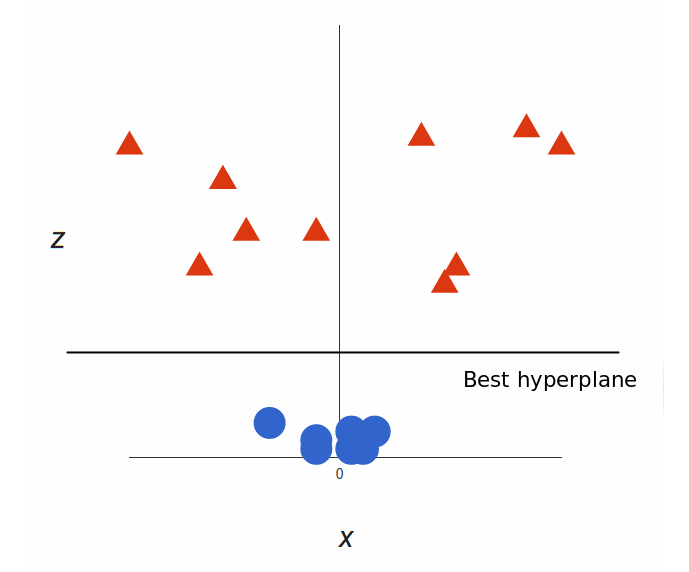
Trong trường hợp trên, dữ liệu huấn luyện khá đơn giản khi có thể phân chia chỉ bằng đường thẳng, ta gọi đây là dữ liệu tuyến tính. Tuy nhiên trong thực tế hầu hết dữ liệu không thể phân chia theo cách đơn giản này



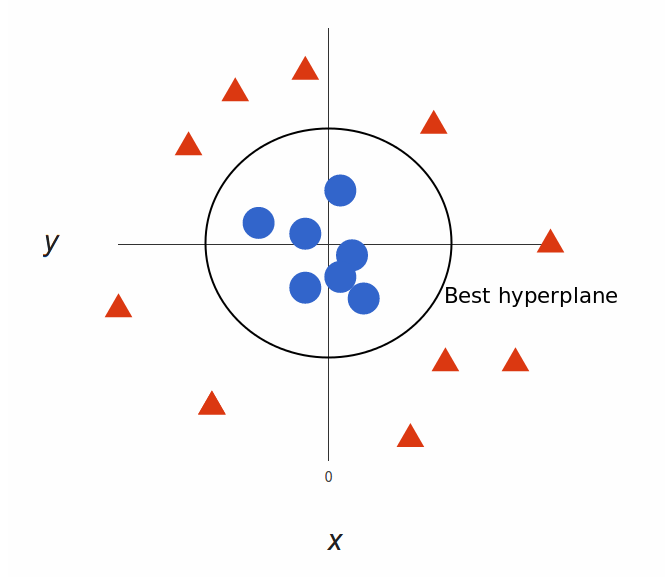
Rõ ràng không thể tìm ra đường phân tách tuyến tính cho trường hợp này. Vì vậy chúng ta sẽ sử dụng thêm một chiều nữa để có một không gian 3 chiều với quy tắc sau z = x2 + y2 . Với quy tắc này, mặt phẳng xz sẽ có dạng như sau:



Từ đây có thể tìm được mặt phẳng phân cực như sau:



Dựa vào quy tắc đã đặt ra ở trên khi trở lại mặt phẳng xy ta sẽ có đường biên có dạng hình tròn như sau:



3. Thủ thuật hạt nhân

Trong ví dụ trên, chúng ta đã tìm thấy một cách để phân loại dữ liệu phi tuyến bằng cách khéo léo ánh xạ không gian đầu vào sang một chiều không gian khác. Tuy nhiên, việc tính toán sự biến đổi này có thể khá tốn kém về mặt tính toán: có thể có rất nhiều chiều phát sinh trong khi mỗi chiều đều cần tính toán phức tạp. Thực hiện việc này cho mọi vectơ trong tập dữ liệu có thể tiêu tốn rất nhiều công sức, vì vậy việc tìm ra một phương pháp đơn giản hơn là cần thiết.

Thật may mắn khi không nhất thiết phải sử dụng toàn bộ các vector, thay vào đó thì chúng ta có thể sử dụng tích vô hướng.

Giả sử như ví dụ trên:

z = x² + y²

Tích vô hướng được phân tích như sau:

a · b = xa · xb + ya · yb + za · zb

a · b = xa · xb + ya · yb + (xa² + ya²) · (xb² + yb²)

Sử dụng SVM dựa trên kết quả phân tích này được gọi là sử dụng hàm hạt nhân. Việc này giúp giảm thiểu các phép tính phức tạp tốn kém khi có thể sử dụng phương pháp phân loại tuyến tính trên các bộ dự liệu phi tuyến tính trong khi không cần đến các phép biến đổi phức tạp. Lưu ý thêm rằng việc sử dụng hàm nhân không phải là một phần của SVM nên hoàn toàn có thể sử dụng nó trong các phương pháp phân loại tuyến tính khác điển hình như hồi quy logic

3. Sử dụng SVM với Phân loại ngôn ngữ tự nhiên?

Sử dụng SVM có thể giúp ta phân loại vector rong không gian nhiều chiều. Áp dụng thuật toán này để phân loại văn bản, điều đầu tiên cần làm là biến một đoạn văn bản thành một vector số để chạy SVM. Phương pháp phổ biến nhất là tính tần số từ, giống như trong Naive Bayes. Điều này có nghĩa là ta coi một văn bản như một bộ từ và đối với mỗi từ xuất hiện trong bộ đó, ta có một đặc điểm của văn bản này. Giá trị của đặc điểm đó sẽ là tần suất xuất hiện của từ đó trong văn bản. Phương pháp này chỉ đơn giản là đếm số lần mỗi từ xuất hiện trong một văn bản và chia nó cho tổng số từ. Vì vậy, trong câu “Tất cả các loài khỉ đều là động vật linh trưởng nhưng không phải tất cả các loài linh trưởng đều là khỉ” từ “khỉ” có tần suất 2/23, và từ “nhưng” có tần suất 1/23.

Để có một giải pháp thay thế nâng cao hơn cho việc tính toán tần số, chúng ta cũng có thể sử dụng TF-IDF. Để thực hiện nó, mọi văn bản trong tập dữ liệu của ta được biểu diễn dưới dạng vector với hàng nghìn (hoặc hàng chục nghìn) chiều, mỗi chiều đại diện cho tần suất xuất hiện của một trong các từ của văn bản. Đây là những gì ta cần phải cung cấp cho SVM để bắt đầu huấn luyện. Ta có thể cải thiện điều này bằng cách sử dụng các kỹ thuật tiền xử lý, như tạo gốc, xóa từ dừng và sử dụng n-gram.

4. Chọn một hàm nhân

Bây giờ chúng ta đã có các vector đặc trưng, ​​việc duy nhất còn lại là chọn một hàm nhân cho mô hình của chúng ta. Mọi vấn đề đều khác nhau, và hàm nhân phụ thuộc vào bộ dữ liệu trông như thế nào. Trong ví dụ trước, dữ liệu được sắp xếp theo các vòng tròn đồng tâm, vì vậy ta đã chọn một nhân phù hợp với các phân bố dữ liệu đó.

Vậy điều gì là tốt nhất để xử lý ngôn ngữ tự nhiên? Liệu có cần một bộ phân loại phi tuyến không hay dữ liệu có thể phân tách tuyến tính? Thực tế chỉ ra rằng tốt nhất nên dựa vào một nhân tuyến tính. Quay lại ví dụ trước, ta có hai đặc trưng trong bộ dữ liệu. Trên thực tế, số đặc trưng có thế lên tới 10 thậm trí hằng trăm đặc trưng khác nhau. Trong khi đó, bộ phân loại NLP sử dụng hàng nghìn đặc trưng, trong khi đó nó chỉ có thể có tối đa một đặc trưng cho mỗi từ xuất hiện trong bộ dữ liệu huấn luyện. Điều này làm thay đổi vấn đề: sử dụng hàm hạt nhân phi tuyến có thể là một ý tưởng tốt trong các trường hợp khác, nhưng việc có qúa nhiều đặc trưng sẽ làm cho hạt nhân phi tuyến trở nên quá khớp với tập dữ liệu. Do đó, phương án tốt nhất là chỉ nên sử dụng nhân tuyến tính kiểu cũ, điều này sẽ mang lại hiệu suất tốt nhất trong những trường hợp này.