Dùng thuật toán Support Vector Machine để giải quyết Data Training và xây dựng Model

1. Vấn đề

Sau khi có một tập các ảnh dữ liệu training là ảnh X-ray về bệnh phổi, và đã kéo được vector ảnh số cho tập dữ liệu này, làm sau để hình thành Model cho bài toán từ tập dữ liệu vector để biết được ảnh nào (bộ vector nào) là có bệnh, và ảnh nào (bộ vector nào) là không có bệnh?

1. Hướng giải quyết

Sau khi chuyển tất cả các ảnh X-ray về dạng vector ảnh số, ta phải tìm ra quy luật phân biệt của các vector ảnh, sau đó dùng thuật toán Support Vector Machine để tìm ra một siêu phẳng để phân chia các ảnh về 2 phía, dựa theo label ta sẽ biết được ảnh nào là có bệnh và ảnh nào là không có bệnh.

Cách giải quyết là phải tìm được một siêu phẳng để phân chia 2 phần có bệnh và không có bệnh sao cho công bằng nhất.

* Giả sử ban đầu ta có bộ dữ liệu training với label có bệnh và không có bệnh có dạng sau:

|  |
| --- |
| () |

Trong đó:

* + Là là vector ảnh số được kéo ra từ ảnh X-ray
  + : Là label với 2 giá trị là có bệnh / không có bệnh. (1/-1 : Tùy theo cách đặt nhãn)
  + : sẽ là mặt phẳng phân chia giữa 2 lớp có bệnh và không có bệnh
  + Vậy với cặp dữ liệu thì khoảng cách từ điểm đó đến mặt phẳng phân chia là:

Với trong đó d là số chiều không gian trên vector ảnh số.

* + Nếu ta tìm được min(|(1)|) thì ta sẽ tìm được khoảng gần nhất từ một điểm trong 2 lớp đến siêu phẳng (margin).
  + Để tối ưu hóa bài toán này ta phải tìm được w và b sao cho margin là lớn nhất.