**Ngày 1/10/2010**

**Nhóm Kiên – Long**

**Using SVM For Time Series Prediction**



# Tổng quan về Support vector regression (SVR)

Ý tưởng cơ bản là ánh xạ dữ liệu x vào không gian đặc trưng có số chiều lớn **Ғ** (high-dimensional feature space) bằng một *ánh xạ phi tuyến* **φ** rồi thực hiện hồi quy tuyến tính trong không gian này:

ƒ(x) = ( ω.φ(x) ) + b với φ : Rn 🡪 Ғ, ω є Ғ, [1.1]

và b là ngưỡng (threshold).

Vì vậy, hồi quy tuyến tính trên không gian có số chiều lớn tương ứng với hồi quy phi tuyến trên không gian đầu vào Rn (không gian có số chiều nhỏ). Nếu ta không thể dùng thủ thuật tính toán cho kernel thì tích vô hướng ω.φ(x) phải được tính trực tiếp trên không gian Ғ này. Khi ánh xạ φ đã được cố định, ta phải tìm ω từ dữ liệu bằng cách tối thiểu hóa tổng của giá trị rủi ro kinh nghiệm (empirical risk) Remp[ƒ] và ||ω2||. Điều này làm tăng tính phẳng(flatness) trong không gian đặc trưng.

Rreg[ƒ] = Remp[ƒ] + λ|| ω2|| = , [1.2]

Với l là kích thước tập mẫu (x1, …, xl), C(.) là hàm chi phí, và λ là hằng tiêu chuẩn (regularization constant). Đối với tập hợp hàm chi phí, đẳng thức [1.2] có thể được tối thiểu hóa bằng cách giải quyết vấn đề cực trị của hàm bậc hai (quadratic programming problem).

Ngoài ra, vector ω có thể được tính thông qua các điểm dữ liệu như sau:

ω = [1.3]

chính là giải pháp cho vấn đề cực trị hàm bậc 2 trình bày ở trên. Ngoài ra, có thể được hiểu một cách trực quan là nó giúp điều chỉnh giá trị tiến gần yi . Từ [1.1] và [1.3] ta có thể suy ra vấn đề trên không gian đầu vào Rn như sau:

ƒ(x) = [1.4]

Trong đó, k(xi, xj) = (φ(xi).φ(xj)) là hàm kernel. Và bất kỳ hàm kernel đối xứng nào cũng phải thỏa định lý Mercer tương ứng với một tích vô hướng trong một không gian đặc trưng nào đó. Thông thường đó là RBF kernel (còn gọi là Gaussian kernel):

k(x,y) = exp(-||x – y||2/ (2σ2))

## Vapnik’s ε-insensitive Loss Function

Đối với hàm chi phí này, những số nhân Lagrange αi, αi\* thường “thưa thớt” tức là chúng khác 0 sau khi tối ưu hóa nếu và chỉ nếu chúng vượt khỏi cận (outside of the boundary). Điều đó có nghĩa là chúng thỏa mãn điều kiện Karush-Kuhn-Tucker. Hàm chi phí ε-insensitive được cho bởi

C(ƒ(x) – y) = [1.5]

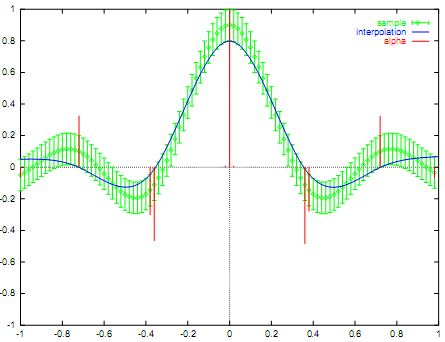
Vấn đề quadratic programming được định ra như sau

Tối thiểu hóa

Với

[1.6]

Càng ít nhiễu, αi, αi\* càng “thưa thớt”. Ngoài ra, nếu ε lớn, chúng sẽ có xu hướng không khớp khi hồi quy, đặc biệt nếu ε quá lớn thì kết quả hồi quy sẽ trở thành hằng.



*Đây là hồi quy trong trường hợp ε-insensitive (dùng kernel B-splines) của hàm sinc. Nó thể hiện tính phẳng tối đa với ε tube được áp xung quanh dữ liệu. Những điểm biên, nơi mà ƒ(xi) – yi = ε sign(αi – αi\*) sẽ được sử dụng để tính ngưỡng b.*

## Huber’s Loss Function

Ưu điểm của hàm này là không đưa thêm bất kỳ độ lệch (bias) nào cũng giống như ε-insensitive. Tuy nhiên, αi và αi\* không còn “thưa thớt” nữa.

C(ƒ(x) – y) = [1.7]

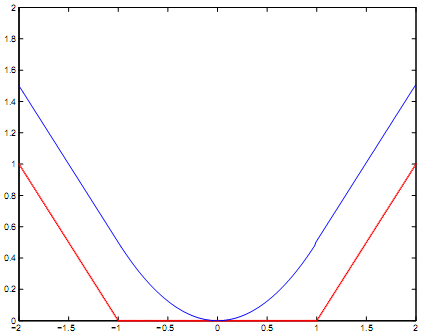
Vấn đề quadratic programming sẽ có dạng

Tối thiểu hóa

Với

[1.8]

Như vậy, về cơ bản, tất cả mẫu đều là support vector.



*ε-insensitive và Huber’s loss function với ε = 1*

## Cách tính ngưỡng b

Phương trình [1.6] và [1.8] dùng để tính αi và αi\*. Để chọn b phù hợp, ta phải sử dụng một cách trực tiếp hơn điều kiện Karush-Kuhn-Tucker để đưa đến vấn đề quadratic programming đã trình bày ở trên. Điểm mấu chốt là chọn được αk, αk\* sao cho độ lỗi δk = ƒ(xk) – yk có thể được xác định một cách độc lập. Trong trường hợp *ε-insensitive,* điều này có nghĩa là chọn những điểm xk trên biên với điều kiện αk, αk\* nằm trong khoảng mở (0;1/𝛌). Trong trường hợp đó, ta được

δk *= ε sign(αi – αi\*)*

Về cơ bản, một giá trị xk đủ để giúp ta xác định b nhưng để vững chắc hơn, ta nên lấy trung bình tất cả những điểm biên với

b = averagek { δk + yk - }

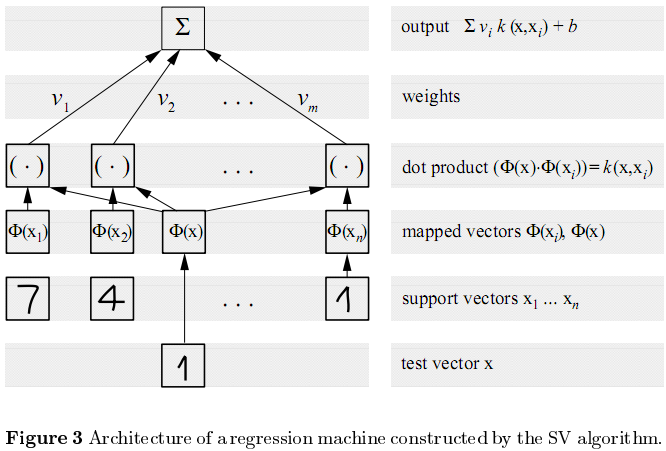
Cho trường hợp của Huber, b được tính tương tự với

δk *=* 𝛌*(αi – αi\*)* với αk, αk\* thuộc về [0; ε/𝛌) nghĩa là những điểm mà phần toàn phương (bậc 2) của hàm chi phí xác định

Chú ý rằng khi ta giải quyết vấn đề quadratic programming với một *bộ tối ưu* để tính toán giá trị double dual ta có thể tìm ra giá trị *biến cơ bản* b.

# Giải thích thêm

## Review các bước thực hiện của SVR trong giai đoạn regression



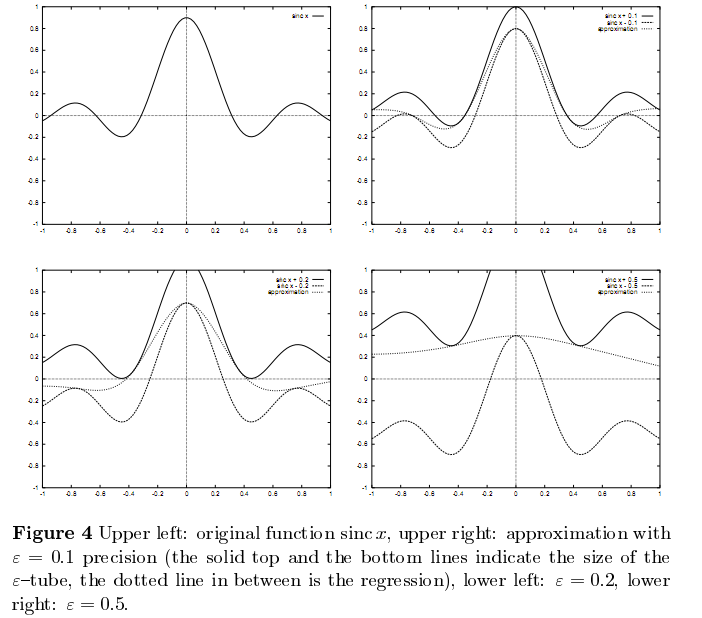
*Cho 1 điểm x trong không gian đầu vào, nhiệm vụ của ta là phải dự đoán giá trị đích ứng với x.*

* Bước 1: Ánh xạ các mẫu từ không gian đầu vào sang không gian đặc trưng bằng ánh xạ .
* Bước 2: Tính tích vô hướng giữa thông qua kernel k(xi, x) với i = 1, 2, …, n.
* Bước 3: Nhân các tích vô hướng này với trọng tương ứng () rồi cộng tất cả lại với nhau.
* Bước 4: Cộng kết quả trên với b sẽ cho ta kết quả dự đoán cuối cùng.

Ghi chú: các giá trị trọng được tính toán trước đó bằng bộ training.

## Giải thích “tính phẳng” (flatness) của hàm

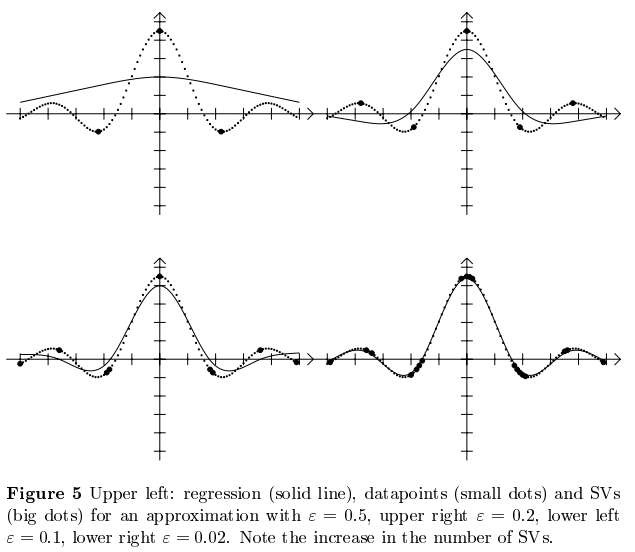
Từ “phẳng” này thực ra xuất phát từ ý nghĩa hình học. Ta hãy ngó qua hình vẽ dưới đây:



Có 3 cái đường, trong đó 2 cái đường trên cùng và dưới cùng tạo thành , còn cái đường ở giữa chính là cái đường hồi qui của ta. “Tính phẳng nhất” ở đây có nghĩa là: trong phạm vi giới hạn bởi cái ống trên, ta phải tìm 1 cái đường sao cho nó ít “uốn lượn” nhất. Ý nghĩa của việc ít “uốn lượn” đơn giản là để tránh overfit.

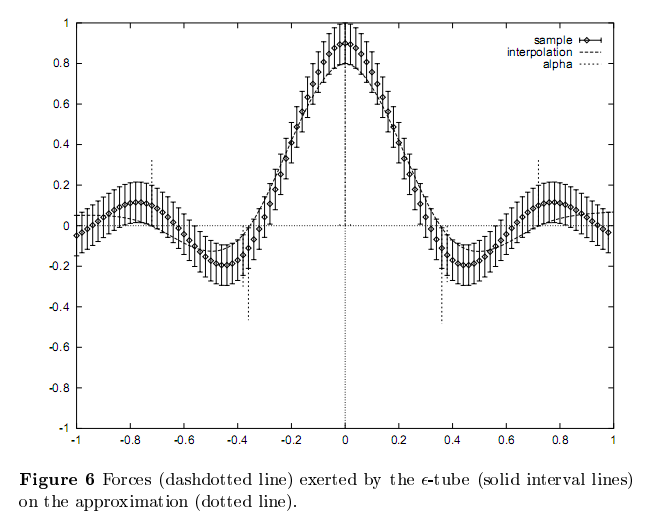
* Thấy ngay nếu quá lớn, cái đường hồi qui của ta có nguy cơ trở thành đường thẳng song song với trục hoành.
* Nếu nhỏ, tiến tới 0 thì cái đường của ta có độ chính xác càng cao. Vậy ta đặt ra cái làm gì cho phức tạp, cứ cho nó = 0 cho độ chính xác nó cao? Ngay dưới đây, ta sẽ có câu trả lời cho chuyện này.

## Mối quan hệ giữa độ chính xác (độ rộng của ) và số điểm SV



Từ hình vẽ, ta nhận thấy ngay cái ống càng rộng (độ chính xác càng thấp) thì số điểm SV cần để cho ra cái đường hồi qui của ta càng ít. Câu trả lời cho câu hỏi phía trên chính là ở chỗ này. Cụ thể là nếu ta cho thì số điểm SV cần để tạo ra đường hồi qui là toàn bộ các điểm train. Thông thường số điểm train này là rất lớn và nó dẫn đến vấn đề tính toán sẽ rất là chậm. Vậy nên, khi áp dụng thực tế ta cần phải cân nhắc giữa độ chính xác và độ phức tạp tính toán.

## Ý nghĩa của nhân tử Lagrange



Nhân tử Lagrange thể hiện cho các lực kéo/ đẩy nhằm đảm bảo cho cái đường hồi qui của ta nằm trong cái ống. Như vậy, chỉ tại những chỗ mà cái đường hồi qui của ta chạm/ vượt ra khỏi cái ống thì ta mới cần đến “lực kéo/ đẩy” Lagrange; những chỗ còn lại, lực này = 0.

# Tối ưu hóa

## Lý thuyết tối ưu hóa

### Tối ưu hóa có ràng buộc (Constrained Optimization)

Xét bài toán tối ưu hóa có ràng buộc sau:

Gọi S là mặt ràng buộc được định bởi g.

Để giải bài toán khó này, ta tìm trước điều kiện ắt:

Giả sử có thêm tại điểm khảo sát có grad g() ≠ 0 (Tại , S có mặt phẳng tiếp xúc có một vector pháp tuyến là grad g().)

Giả sử tại này hàm số f đã có cực trị có ràng buộc rồi cho một đường C khả vi, liên tục /S qua : hàm vector C có C’ liên tục trên khoảng I và

Hàm hợp f[C(t)] là hàm số theo một biến số có cực trị tại t0 (vì theo giả thiết f đã có cực trị tại X0 rồi.)

(Theo qui tắc mắt xích)

Như vậy tại giải pháp, vector gradient của hàm mục tiêu f phải vuông góc với mặt ràng buộc (feasible set) được định nghĩa bởi g(x) = 0.

hay

Ở đây, được gọi là nhân tử Lagrange.

### Ràng buộc là bất đẳng thức

Xét bài toán tối ưu hóa có ràng buộc là bất đẳng thức như sau:

Ta xét 2 trường hợp:

* Tại giải pháp, g(x)=0. Có ngay phải tồn tại nhân tử Lagrange để:

[1]

và hơn nữa, >0 (Nếu không, ta sẽ có thể giảm f(x) đến vô hạn theo chiều - mà vẫn nằm trong miền ràng buộc.)

* Tại giải pháp, g(x)>0. Hiểu ngay ta đang đứng tại cực tiểu của f. Do đó:

và

Tổng hợp lại, ta có điều kiện KKT (Karush–Kuhn–Tucker) cho bài toán tối ưu hóa có ràng buộc là bất đẳng thức:

### Convex Optimization

Giả sử rằng f(x) convex, g(x) concave và cả hai là khả vi liên tục.

Đặt:

Từ [1] ta có:

Cố định , L convex theo x và có một cực tiểu duy nhất.

Cố định x, L tuyến tính theo .

Đặt:

có một cực đại duy nhất ()

* Nếu cực đại này xảy ra tại : , nghĩa là ta đang đứng tại cực tiểu toàn cục của f.
* Nếu cực đại này xảy ra tại , ta đang ở biên của miển ràng buộc.

Tóm lại, bài toán:

tương đương với:

Đây được gọi là vấn đề đôi (dual problem), và giải pháp của nó cũng là giải pháp của vấn đề ban đầu (primal problem):

## Vấn đề tối ưu hóa trong SVM

### Hàm rủi ro

Ta có bộ train

Giả sử các điểm này là IID (independent and identically distributed) và được phát sinh bởi hàm phân phối P(x, y). Mục đích của ta là tìm một hàm f sao cho hàm rủi ro R đạt cực tiểu trên tập X:

Trong đó, c(x, y, f(x)) làm hàm chi phí, thể hiện mức phạt của ta đối với lỗi.

Nếu ta không biết P(x, y), ta có thể tính R[f] như sau:

Hàm rủi ro trên được gọi là hàm rủi ro kinh nghiệm (empirical risk function).

Tuy nhiên, để tránh vấn đề overfit, ta có hàm rủi ro chuẩn hóa (regularized risk function) như sau:

với được gọi là hằng chuẩn hóa.

### Hàm chi phí

### Biểu diễn lại vấn đề tối ưu hóa theo nhân tử Lagrange

Từ hàm rủi ro chuẩn hóa, ta thể hiện lại vấn đề như sau:

Tìm min của:

với ràng buộc

Từ đây, ta có dạng dual như sau:

Tìm max của:

trong đó:

với ràng buộc:

