

Học Máy

(Machine Learning)

Thân Quang Khoát

khoattq@soict.hust.edu.vn

Viện Công nghệ thông tin và Truyền thông
Trường Đại học Bách Khoa Hà Nội
Năm 2015

Nội dung môn học:

- Giới thiệu chung
- **Các phương pháp học có giám sát**
 - **Hồi quy tuyến tính (Linear regression)**
- Đánh giá hiệu năng hệ thống học máy
- Các phương pháp học không giám sát

Hồi quy tuyến tính: Giới thiệu

- **Bài toán hồi quy:** cần học một hàm $y = f(\mathbf{x})$ từ một tập học cho trước $\mathbf{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M)\}$, trong đó $y_i \approx f(\mathbf{x}_i) \in R$, $\mathbf{x}_i = (x_{i1}, \dots, x_{in})^T \in R^n$.
(nét đậm ký hiệu véctor)
- **Mô hình hồi quy tuyến tính:** nếu hàm $y = f(\mathbf{x})$ được giả thuyết là hàm tuyến tính.

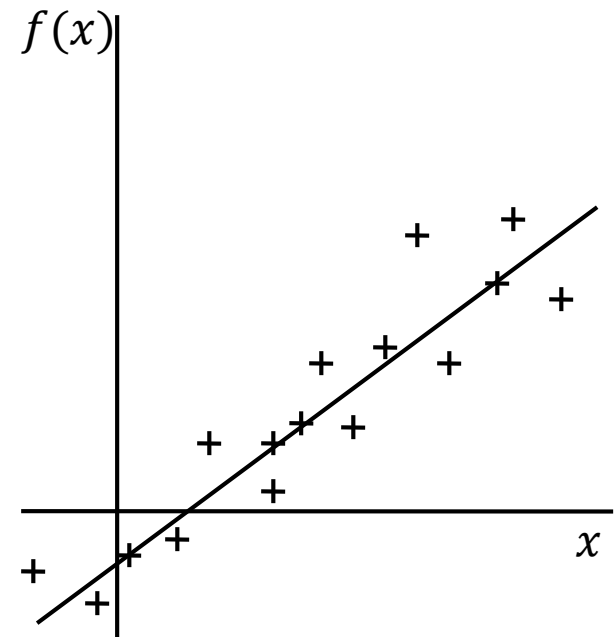
$$f(\mathbf{x}) = w_0 + w_1 x_1 + \dots + w_n x_n$$

- Các $w_i \in R$ là hệ số của hàm hồi quy.
- Học một hàm hồi quy tuyến tính thì tương đương với việc học véctor trọng số $\mathbf{w} = (w_0, w_1, \dots, w_n)^T$.

Hồi quy tuyến tính: Ví dụ

Hàm tuyến tính $f(x)$ nào phù hợp?

0.13	-0.91
1.02	-0.17
3.17	1.61
-2.76	-3.31
1.44	0.18
5.28	3.36
-1.74	-2.46
7.93	5.56
...	...



Ví dụ: $f(x) = -1.02 + 0.83x$

Phán đoán tương lai

- Đối với mỗi quan sát $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, trong đó $x_i \in R$

- Giá trị **đầu ra mong muốn** $c_x \in R$
(Không biết trước đối với các quan sát trong tương lai)

- Giá trị **phán đoán** (bởi hệ thống)

$$y_x = w_0 + w_1 x_1 + \dots + w_n x_n$$

→ Ta thường mong muốn y_x xấp xỉ tốt c_x

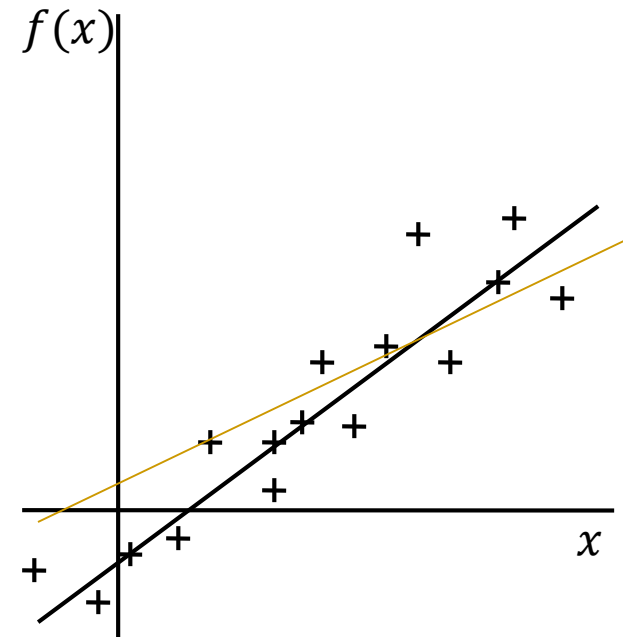
- **Phán đoán cho quan sát tương lai** $\mathbf{z} = (z_1, z_2, \dots, z_n)^T$

- Cần dự đoán giá trị đầu ra, bằng cách áp dụng hàm mục tiêu đã học được f :

$$f(\mathbf{z}) = w_0 + w_1 z_1 + \dots + w_n z_n$$

Học hàm hồi quy

- **Mục tiêu học:** học một hàm f^* sao cho khả năng phán đoán trong tương lai là tốt nhất.
 - Tức là sai số $c_z - f^*(z)$ là nhỏ nhất cho các quan sát tương lai z .
 - Khả năng tổng quát hóa (generalization) là tốt nhất.
- **Vấn đề:** Có vô hạn hàm tuyến tính!!
 - Làm sao để học? Quy tắc nào?
- Dùng một tiêu chuẩn để đánh giá.
 - Tiêu chuẩn thường dùng là hàm lỗi (generalization error, loss function, ...)



Hàm đánh giá lỗi

- Định nghĩa hàm lỗi E

- Lỗi của hệ thống đối với mỗi quan sát $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$:

$$r(\mathbf{x}) = [c_x - f^*(\mathbf{x})]^2 = (c_x - w_0 - w_1 x_1 - \dots - w_n x_n)^2$$

- Lỗi của hệ thống đối với toàn bộ không gian của \mathbf{x} :

$$E = E_x[r(\mathbf{x})] = E_x[(c_x - f^*(\mathbf{x}))^2]$$

- Mục tiêu học là tìm hàm f^* mà E là nhỏ nhất.

$$f^* = \arg \min_{f \in H} E_x[r(\mathbf{x})]$$

- Trong đó H là không gian của hàm f .

- Tuy nhiên, trong quá trình học ta không thể làm việc được với bài toán này.

Hàm lỗi thực nghiệm

- Ta chỉ quan sát được một tập $\mathbf{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M)\}$. Cần học hàm f từ đó.
- **Lỗi thực nghiệm** (empirical loss; residual sum of squares)
 - Là một xấp xỉ của $E_{\mathbf{x}}[r(\mathbf{x})]$ trên tập học \mathbf{D}

□ Định nghĩa:

$$RSS(f) = \sum_{i=1}^M (y_i - f(\mathbf{x}_i))^2 = \sum_{i=1}^M (y_i - w_0 - w_1 x_{i1} - \dots - w_n x_{in})^2$$

- Nhiều phương pháp học thường gắn với RSS .

Bình phương tối thiểu

- Cho trước \mathbf{D} , ta đi tìm hàm f mà có RSS nhỏ nhất.

$$f^* = \arg \min_{f \in H} RSS(f)$$

$$\Leftrightarrow \mathbf{w}^* = \arg \min_{\mathbf{w}} \sum_{i=1}^M (y_i - w_0 - w_1 x_{i1} - \cdots - w_n x_{in})^2 \quad (1)$$

- Đây được gọi là **bình phương tối thiểu** (least squares).
- Tìm nghiệm \mathbf{w}^* bằng cách lấy đạo hàm của RSS và giải phương trình $RSS' = 0$. Thu được:

$$\mathbf{w}^* = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$$

- Trong đó \mathbf{A} là ma trận dữ liệu cỡ $M \times (n + 1)$ mà hàng thứ i là $(1, x_{i1}, \dots, x_{in})$; \mathbf{B}^{-1} là ma trận nghịch đảo; $\mathbf{y} = (y_1, \dots, y_M)^T$.
- **Chú ý: giả thuyết $\mathbf{A}^T \mathbf{A}$ tồn tại nghịch đảo.**

Bình phương tối thiểu: thuật toán

- Input: $\mathbf{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M)\}$
- Output: \mathbf{w}^*
- Học \mathbf{w}^* bằng cách tính:

$$\mathbf{w}^* = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{y}$$

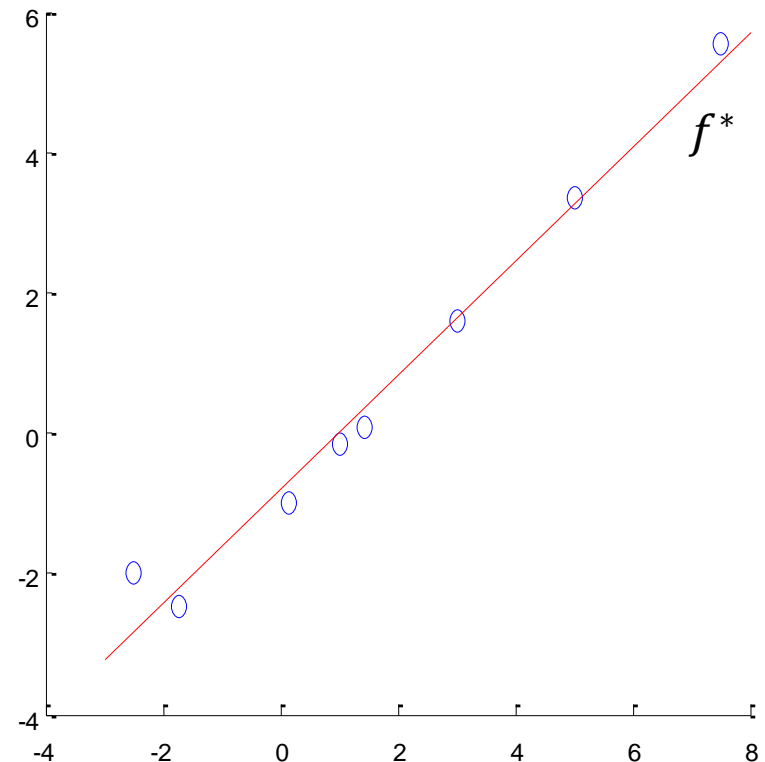
- Trong đó \mathbf{A} là ma trận dữ liệu cỡ $M \times (n + 1)$ mà hàng thứ i là một véctơ $\mathbf{A}_i = (1, x_{i1}, \dots, x_{in})$, \mathbf{B}^{-1} là ma trận nghịch đảo, $\mathbf{y} = (y_1, \dots, y_M)^T$.
- Chú ý: giả thuyết $\mathbf{A}^T \mathbf{A}$ tồn tại nghịch đảo.
- Phán đoán cho quan sát mới \mathbf{x} :
$$y_x = w_0^* + w_1^* x_1 + \dots + w_n^* x_n$$

Bình phương tối thiểu: ví dụ

Kết quả học bằng bình phương tối thiểu

0.13	-1
1.02	-0.17
3	1.61
-2.5	-2
1.44	0.1
5	3.36
-1.74	-2.46
7.5	5.56

$$f^*(x) = 0.81x - 0.78$$



Bình phương tối thiểu: nhược điểm

- Nếu $A^T A$ không tồn tại nghịch đảo thì không học được.
 - Nếu các thuộc tính (cột của A) có phụ thuộc lẫn nhau.
- Độ phức tạp tính toán lớn do phải tính ma trận nghịch đảo.
→ Không làm việc được nếu số chiều n lớn.
- Khả năng overfitting cao vì việc học hàm f chỉ quan tâm tối thiểu lỗi đối với tập học đang có.

Ridge regression (1)

- Cho trước $\mathbf{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M)\}$, ta đi giải bài toán:

$$\begin{aligned} f^* &= \arg \min_{f \in H} \text{RSS}(f) + \lambda \|\mathbf{w}\|_2^2 \\ \Leftrightarrow \mathbf{w}^* &= \arg \min_{\mathbf{w}} \sum_{i=1}^M (y_i - \mathbf{A}_i \mathbf{w})^2 + \lambda \sum_{j=0}^n w_j^2 \end{aligned} \quad (2)$$

- Trong đó $\lambda > 0$ là một hằng số phạt.
- $\mathbf{A}_i = (1, x_{i1}, \dots, x_{in})$ tương ứng với quan sát x_i
- Đại lượng chuẩn tắc (phạt) $\lambda \|\mathbf{w}\|_2^2$:
 - Có vai trò hạn chế độ lớn của \mathbf{w}^* (hạn chế không gian hàm f).
 - Đánh đổi chất lượng của hàm f đối với tập học \mathbf{D} , để có khả năng phán đoán tốt hơn với quan sát tương lai.

Ridge regression (2)

- Tìm nghiệm \mathbf{w}^* bằng cách lấy đạo hàm của RSS và giải phương trình $RSS' = 0$. Thu được:
$$\mathbf{w}^* = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}_{n+1})^{-1} \mathbf{A}^T \mathbf{y}$$
 - Trong đó \mathbf{A} là ma trận dữ liệu cỡ $M \times (n + 1)$ mà hàng thứ i là $(1, x_{i1}, \dots, x_{in})$; $\mathbf{y} = (y_1, \dots, y_M)^T$; \mathbf{I}_{n+1} là ma trận đơn vị cỡ $n + 1$.
- So sánh với phương pháp bình phương tối thiểu:
 - Tránh được trường hợp ma trận dữ liệu suy biến. Hồi quy Ridge luôn làm việc được.
 - Khả năng overfitting thường ít hơn.
 - Lỗi trên tập học có thể nhiều hơn.
- **Chú ý:** chất lượng của phương pháp phụ thuộc rất nhiều vào sự lựa chọn của tham số λ .

Ridge regression: thuật toán

- Input: $\mathbf{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_M, y_M)\}$, hằng số $\lambda > 0$
- Output: \mathbf{w}^*
- Học \mathbf{w}^* bằng cách tính:

$$\mathbf{w}^* = (\mathbf{A}^T \mathbf{A} + \lambda \mathbf{I}_{n+1})^{-1} \mathbf{A}^T \mathbf{y}$$

- Phán đoán cho quan sát mới \mathbf{x} :

$$y_x = w_0^* + w_1^* x_1 + \dots + w_n^* x_n$$

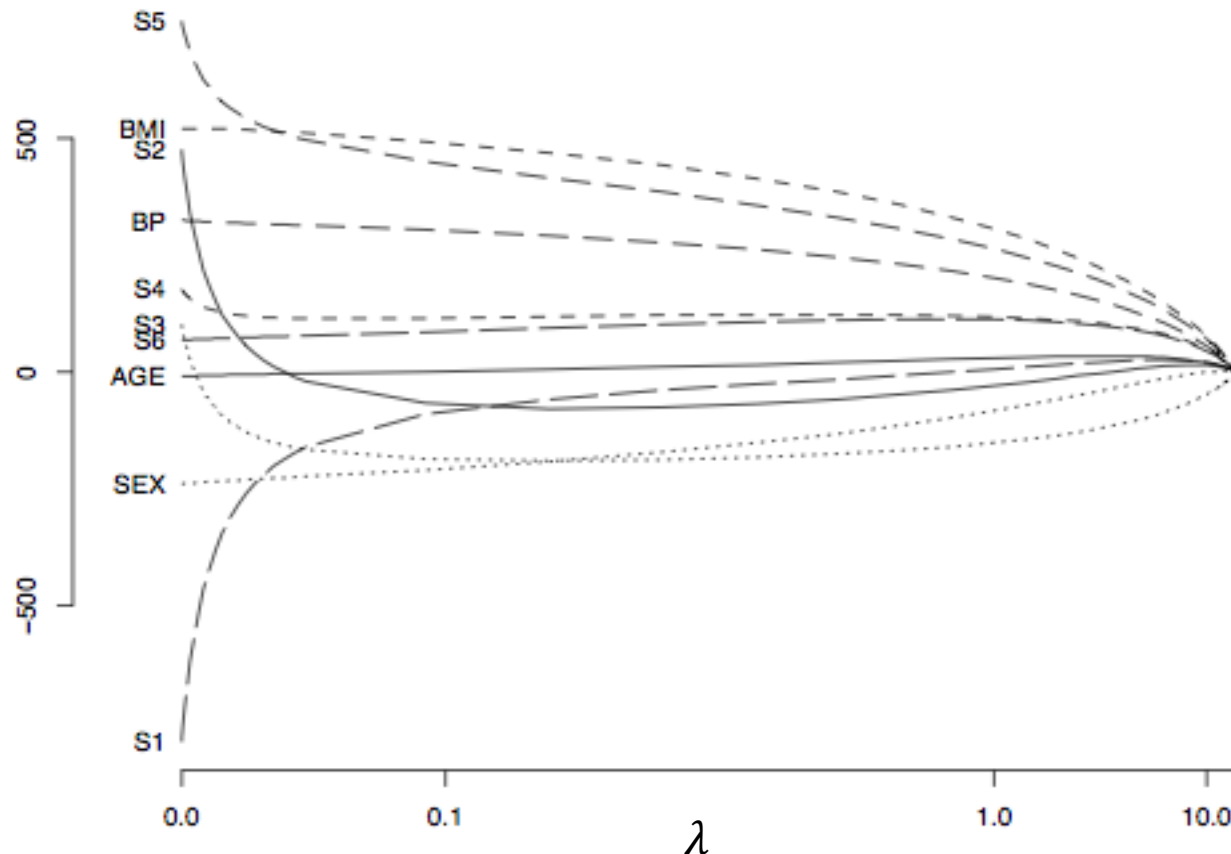
Ridge regression: ví dụ

- Xét tập dữ liệu Prostate gồm 67 quan sát dùng để học, và 31 quan sát dùng để kiểm thử. Dữ liệu gồm 8 thuộc tính.

w	Least squares	Ridge
0	2.465	2.452
lcavol	0.680	0.420
lweight	0.263	0.238
age	-0.141	-0.152
lbph	0.210	0.002
svi	0.305	0.094
lcp	-0.288	-0.051
gleason	-0.021	0.232
pgg45	0.267	-0.056
Test RSS	0.521	0.492

Ridge regression: ảnh hưởng của λ

- Vector $\mathbf{w}^* = (w_0, S1, S2, S3, S4, S5, S6, AGE, SEX, BMI, BP)$ thay đổi khi hệ số phạt λ thay đổi.



Câu hỏi ôn tập

- Viết chi tiết từng bước giải để tìm nghiệm cho bài toán (1) và (2).
- Tìm nghiệm của bài toán (2) khi loại bỏ w_0 ra khỏi đại lượng phạt.