Seminar ngày 5/10/09

MÔ HÌNH HỒI QUI TUYẾN TÍNH

VÕ ĐÌNH BẢY

Hồi qui tuyến tính

Mục tiêu của hồi qui là tiên đoán giá trị của một hay nhiều biến (liên tục) mục tiêu *t* khi cho trước giá trị của vector D-chiều x.

Đơn giản nhất là sử dụng công thức dạng y = ax + b.

Công thức hồi qui đơn giản (1)

Công thức: $y = ax + b$.

Khi ấy, với $X = \{x_1, x_2, \dots x_N\}$ và $T = \{t_1, t_2, \dots, t_N\}$. Ta có thể tìm công thức hồi qui như sau:

Xét hàm lỗi SE $= \sum_{i=1}^{N}(t_i - y_i) = \sum_{i=1}^{N}[t_i - (ax_i + b)]$

Cực tiểu hàm lỗi để nhận được các hệ số a, b.

Công thức hồi qui đơn giản (2)

Cực tiểu hàm lỗi để nhận được các hệ số a, b.

Ta có:

$$\frac{\partial SE}{da} = -2\sum_{i=1}^{N}[t_i - (ax_i + b)]x_i$$

$$\frac{\partial SE}{db} = -2\sum_{i=1}^{N}[t_i - (ax_i + b)]$$

Công thức hồi qui đơn giản (3)

Giải hệ trên với biến là a, b:

$$\begin{cases} \sum_{i=1}^{N}[t_i-(ax_i+b)]x_i=0 \\ \sum_{i=1}^{N}[t_i-(ax_i+b)]=0 \end{cases} \Rightarrow \begin{cases} a=\dfrac{\sum_{i=1}^{N}(x_i-mean_X)(t_i-mean_T)}{\sum_{i=1}^{N}(x_i-mean_X)^2} \\ b=\dfrac{\sum_{i=1}^{N}t_i}{N}-a\dfrac{\sum_{i=1}^{N}x_i}{N}=mean_T-a*mean_X \end{cases}$$

Trong đó $mean_X$, $mean_T$ là giá trị trung bình của X và T.

Công thức hồi qui đơn giản (4)

Ví dụ:

X	T
0	0.3
0.2	0.8
0.5	1
0.6	0.9
1	0.01

$$\Rightarrow \begin{cases} a = \dfrac{\sum_{i=1}^{N}(x_i - mean_X)(t_i - mean_T)}{\sum_{i=1}^{N}(x_i - mean_X)^2} = -0.295 \\ b = \dfrac{\sum_{i=1}^{N} t_i}{N} - a\dfrac{\sum_{i=1}^{N} x_i}{N} = mean_T - a * mean_X = 0.738 \end{cases}$$

Hay phương trình là: y = -0.295x+0.738!

Công thức hồi qui đơn giản (4)

X	T
0	0.3
0.2	0.8
0.5	1
0.6	0.9
1	0.01

Hàm dự đoán: y = -0.295x+0.738

X	T’
0.1	0.71
0.8	0.5

$$\Leftarrow \begin{cases} a = -0.295 \\ b = 0.738 \end{cases}$$

Dạng đơn giản – Đa thức

Hồi qui tuyến tính cơ sở (1)

Một cách khác là sử dụng đường cong đa thức:

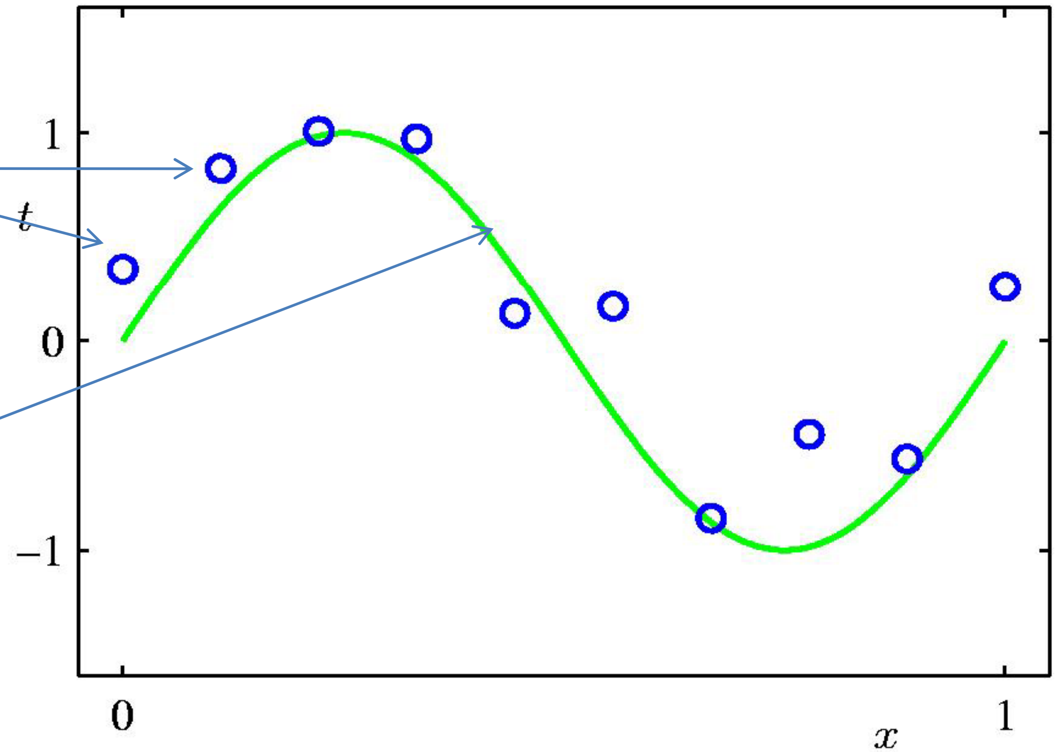
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Tùy theo giá trị M, chúng ta có hàm xấp xỉ với các giá trị (x_i, t_i) được cho.

Hàm hồi qui tuyến tính cơ sở (2)

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Các điểm dữ liệu (x_i, t_i)



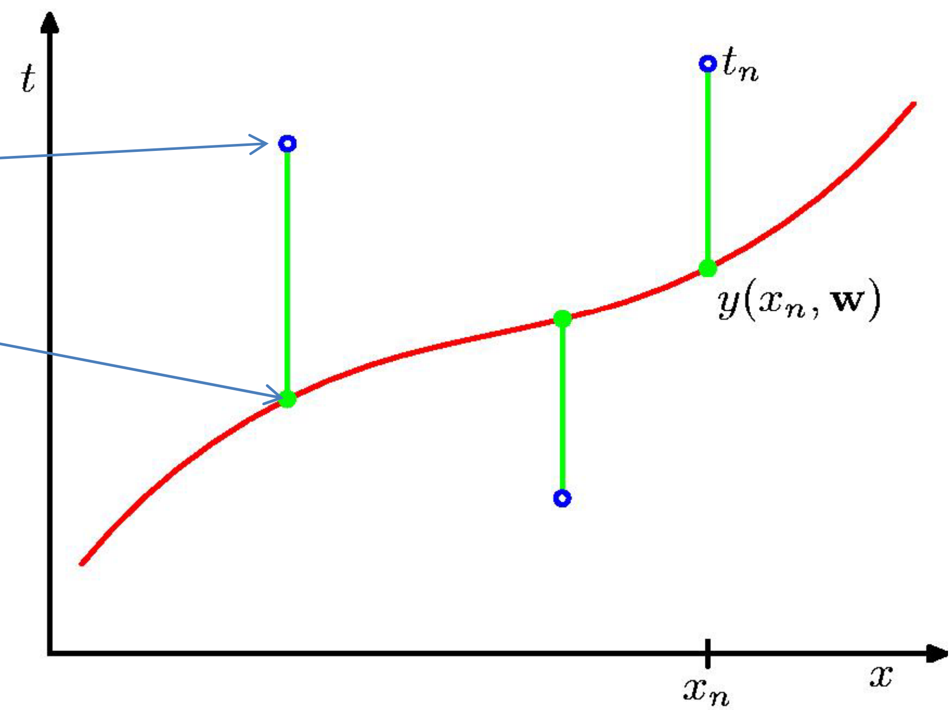
Hàm cần dự đoán ⇒ Cần xác định w_0, …, w_M.

Hàm lỗi (Sum-of-Squares Error Function)

t thực tế

Giá trị ước lượng

Lỗi: y(x,w) - t



$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

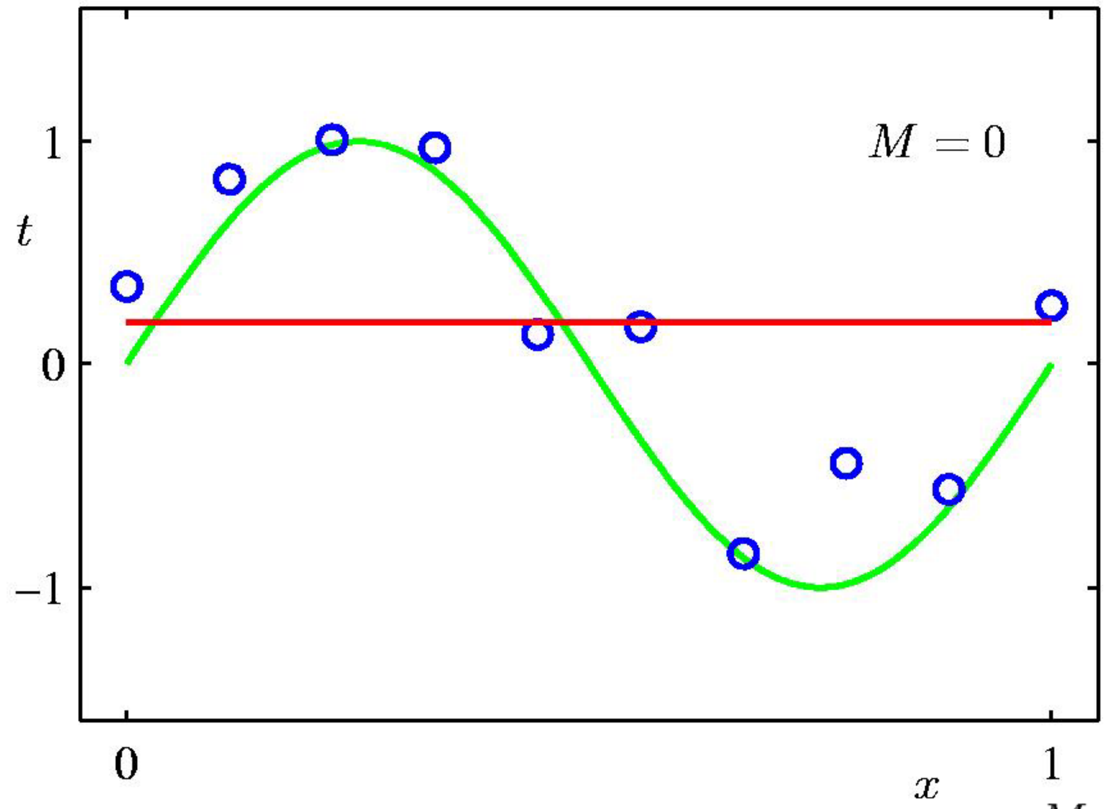
Hàm lỗi (2)

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2$$

Tìm w sao cho E(w) đạt min

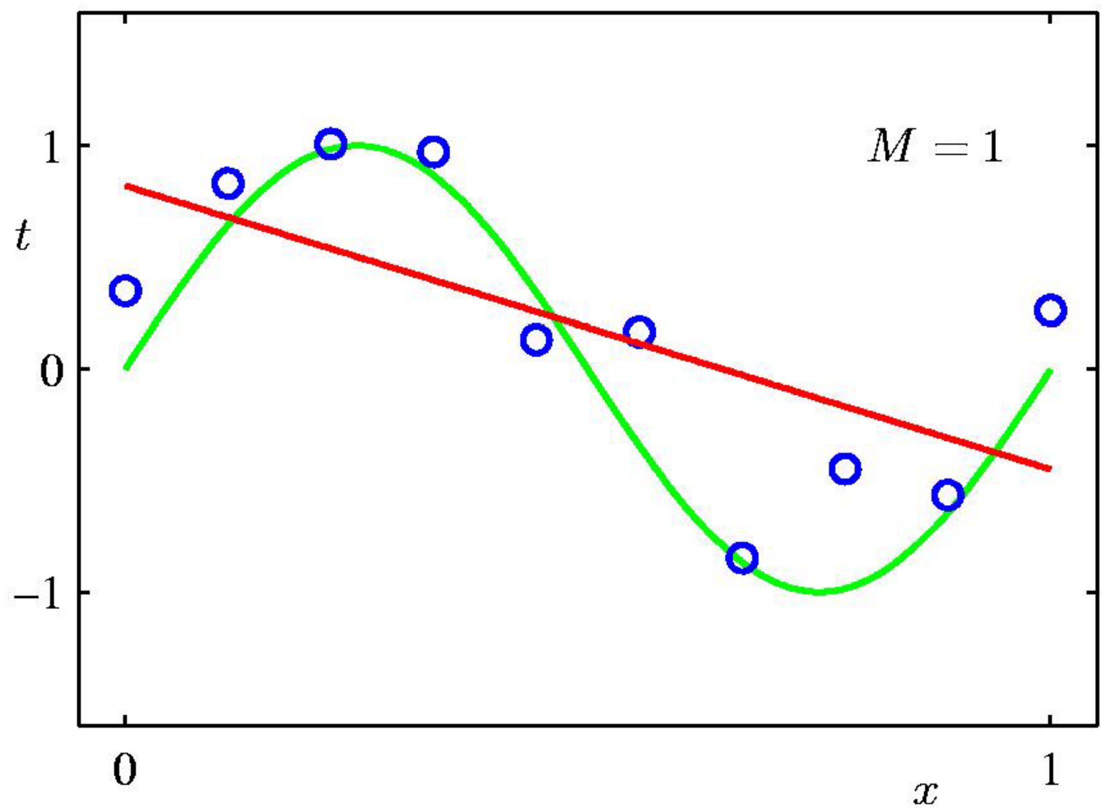
$\Rightarrow$ Giải bài toán cực trị hàm nhiều biến

Hàm xấp xỉ với M = 0



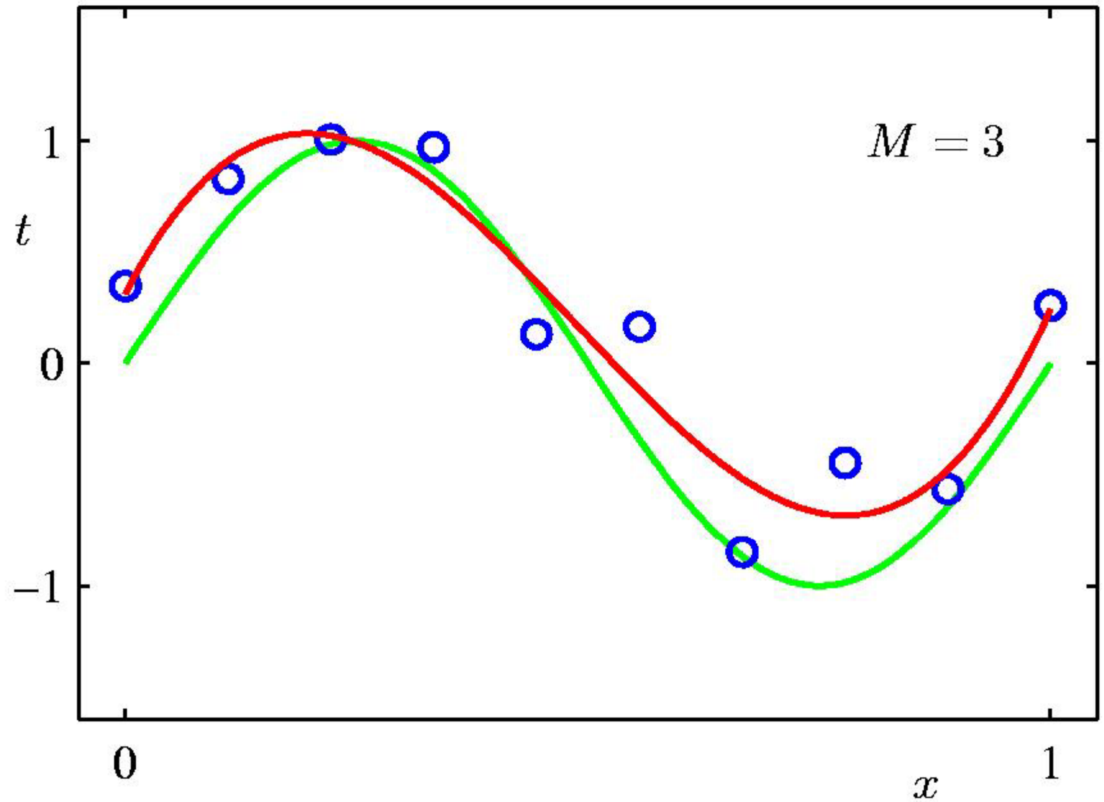
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Hàm xấp xỉ với M = 1



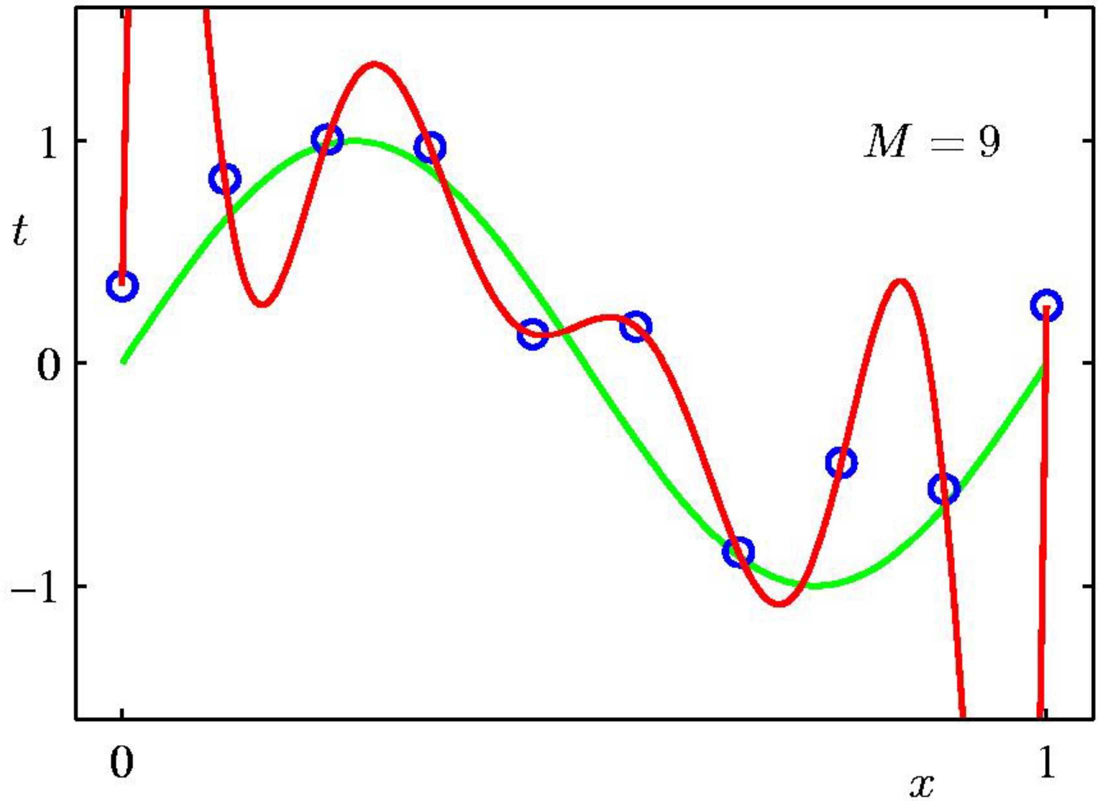
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Hàm xấp xỉ với M = 3



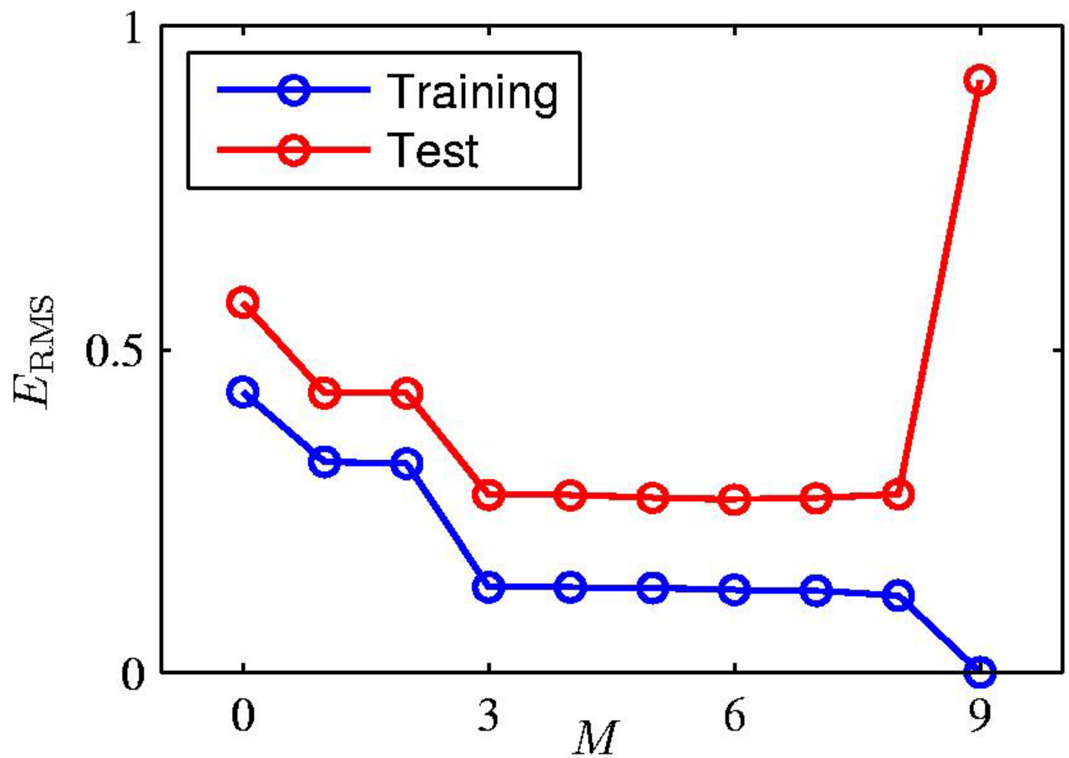
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Hàm xấp xỉ với M = 9



$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

Over-fitting



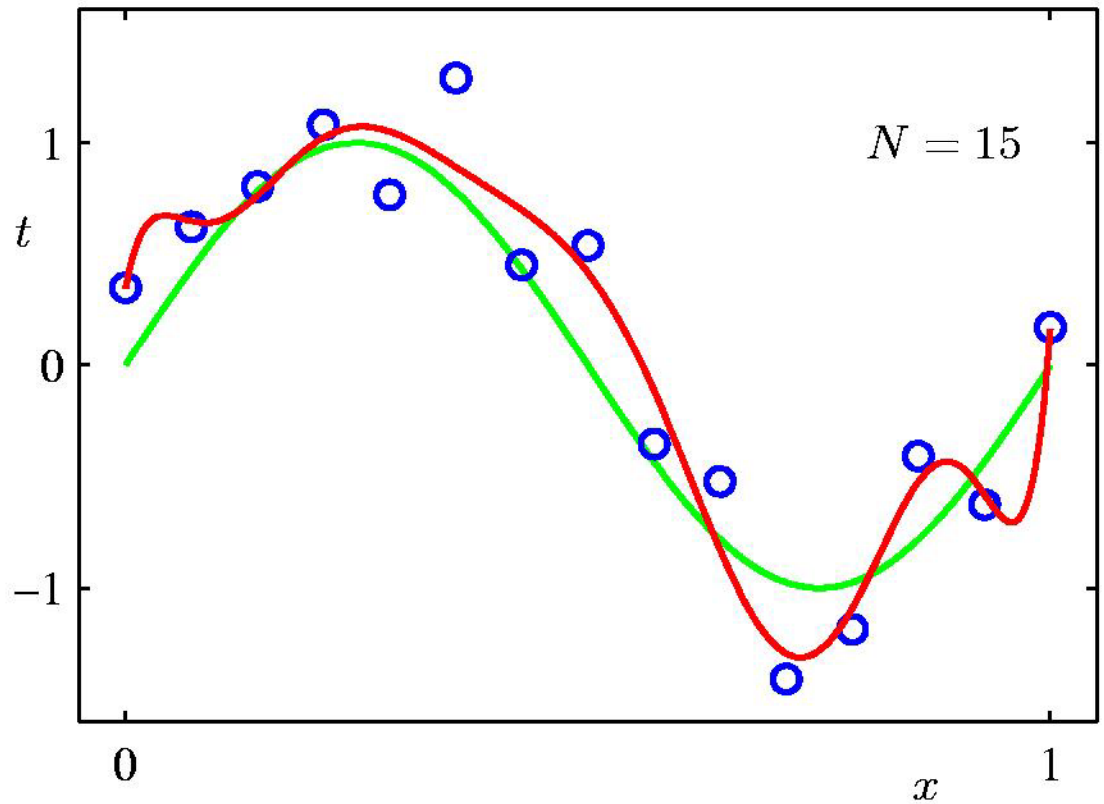
Root-Mean-Square (RMS) Error: $E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^\star)/N}$

Các hệ số tương ứng với M

	$M=0$	$M=1$	$M=3$	$M=9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

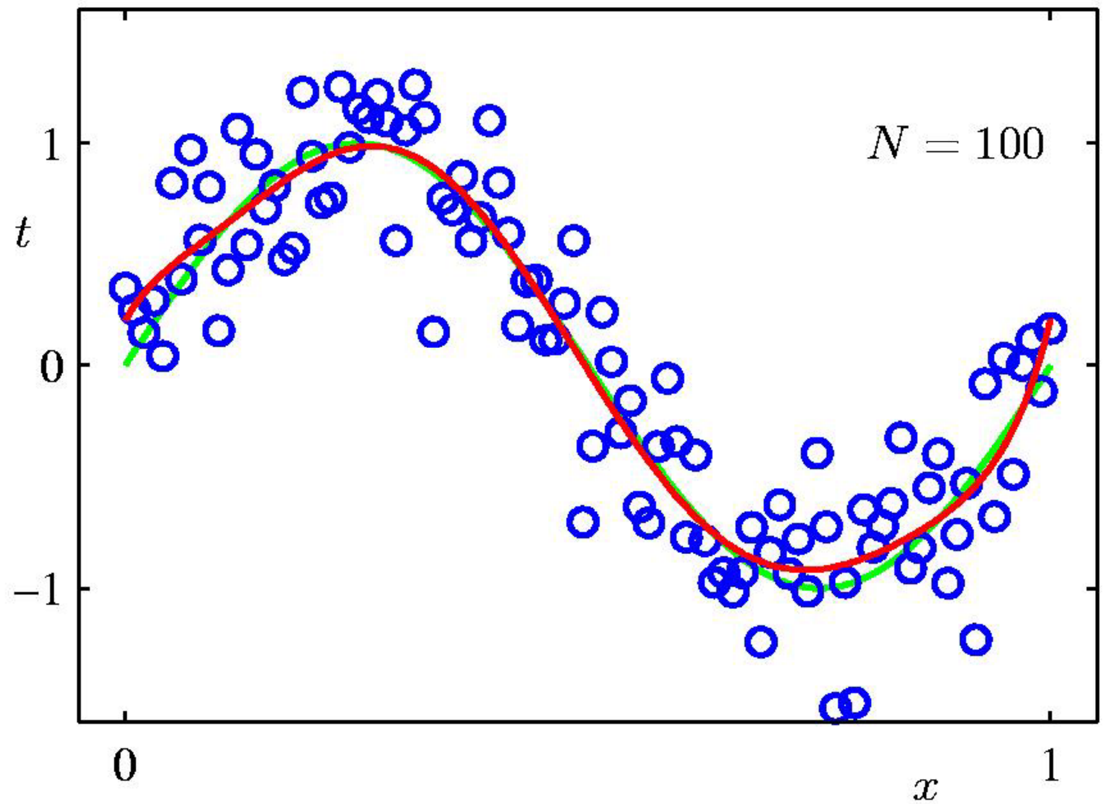
Kích thước dữ liệu: $N = 15$

Hàm xấp xỉ với M = 9



Kích thước dữ liệu: $N = 100$

Hàm xấp xỉ với M = 9



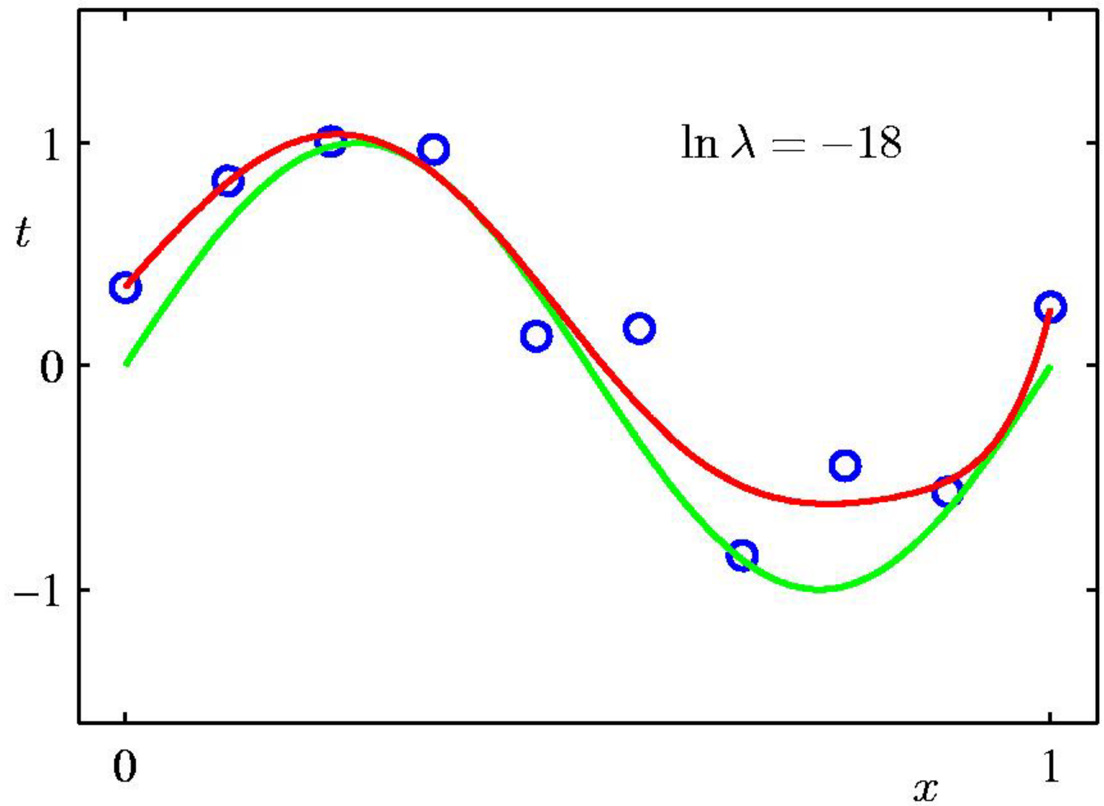
Mở rộng công thức hàm lỗi

Thêm hàm phạt (theo λ và w)

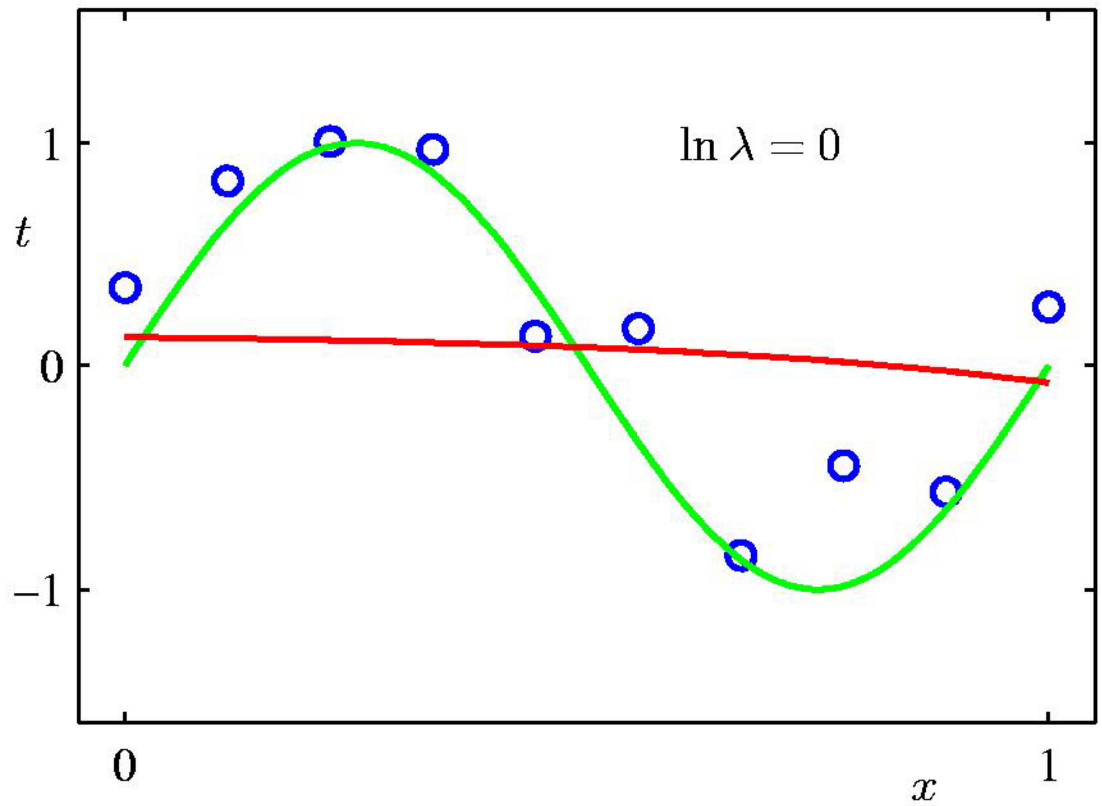
$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Ngoài w, cần chọn λ phù hợp để lỗi đạt được là min.

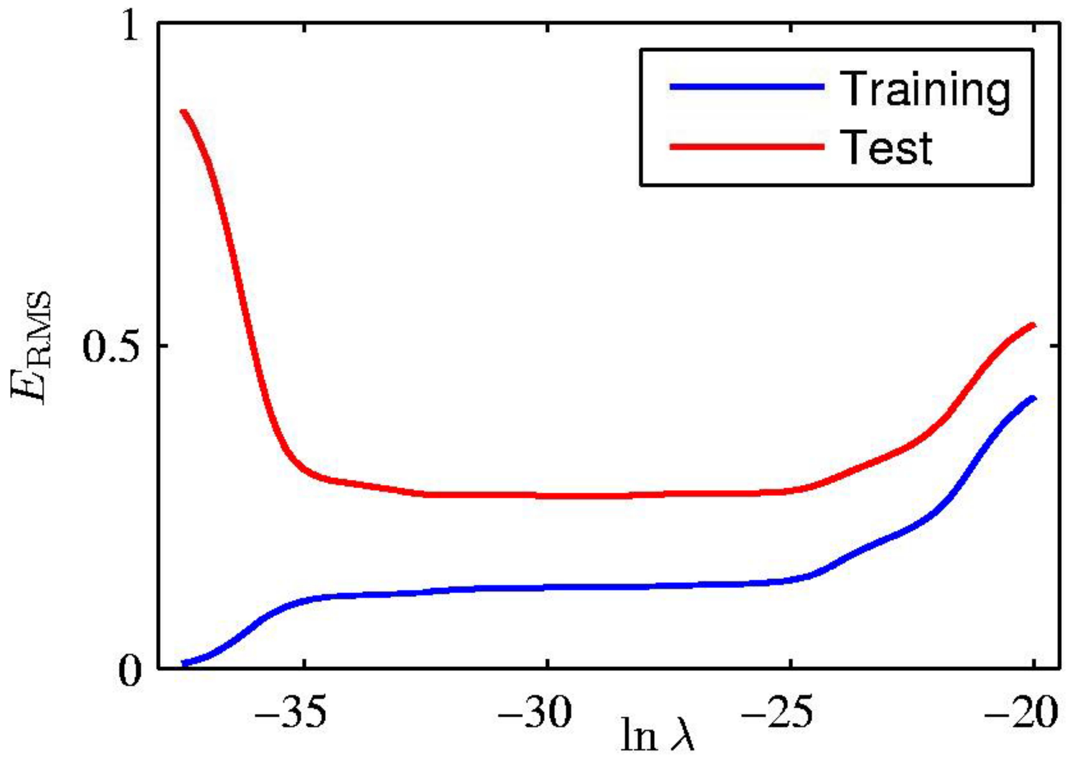
Hệ số λ: $\ln \lambda = -18$



Hệ số λ: $\ln\lambda = 0$



Lỗi với hệ số λ: E_{RMS} với $\ln \lambda$



Các hệ số tương ứng với λ

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

Mở rộng hàm

Hàm hồi qui tuyến tính cơ sở (1)

Công thức tổng quát:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Trong đó $\phi_j(x)$ là các hàm cơ sở (basis functions).

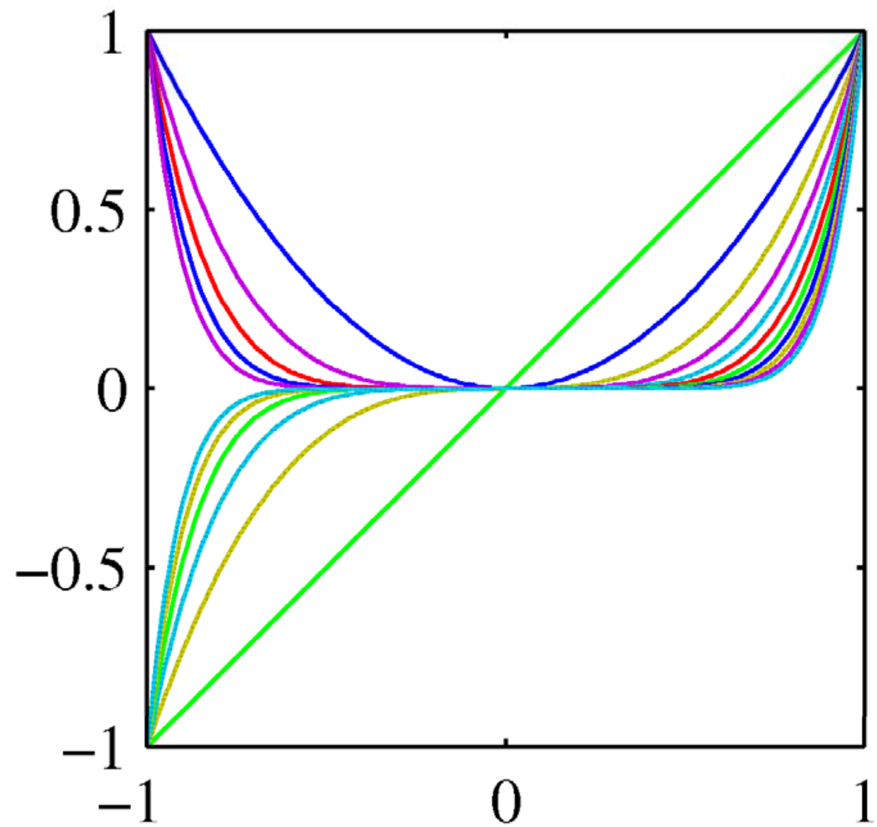
$w = (w_0, w_1, \ldots, w_{M-1})^T$ và $\phi = (\phi_0, \phi_1, \ldots, \phi_{M-1})^T$.

Hàm hồi qui tuyến tính cơ sở (2)

Hàm cơ sở dạng đa thức:

$$\phi_j(x) = x^j.$$

≡ Hàm cơ bản dạng đa thức



Hàm hồi qui tuyến tính cơ sở (3)

Hàm cơ sở dạng Gaussian:

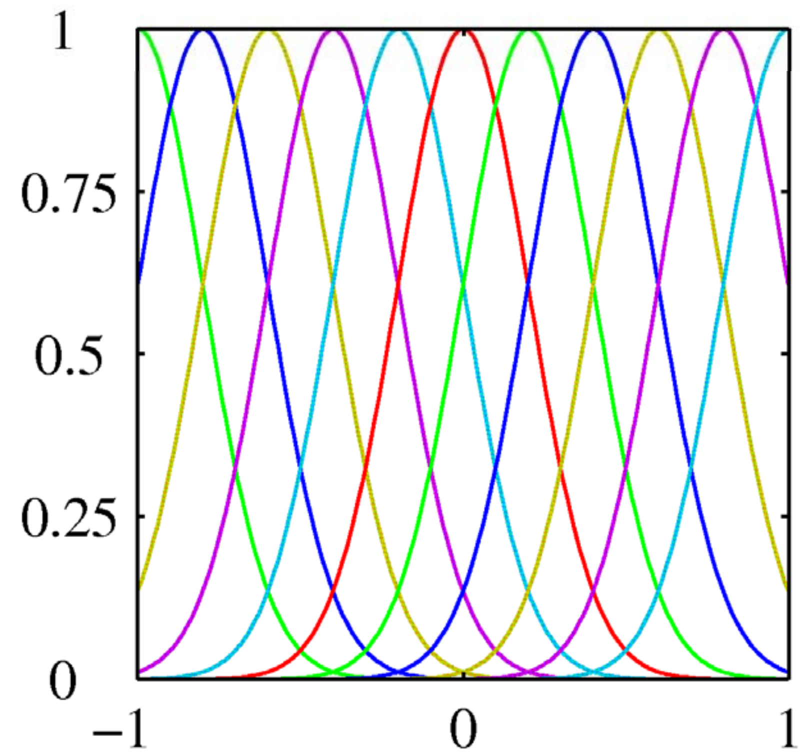
$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

Trong đó μ_j được tính theo công thức:

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu, \sigma^2\right) x \, \mathrm{d}x = \mu$$

Hoặc:

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu \quad \text{với} \quad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$



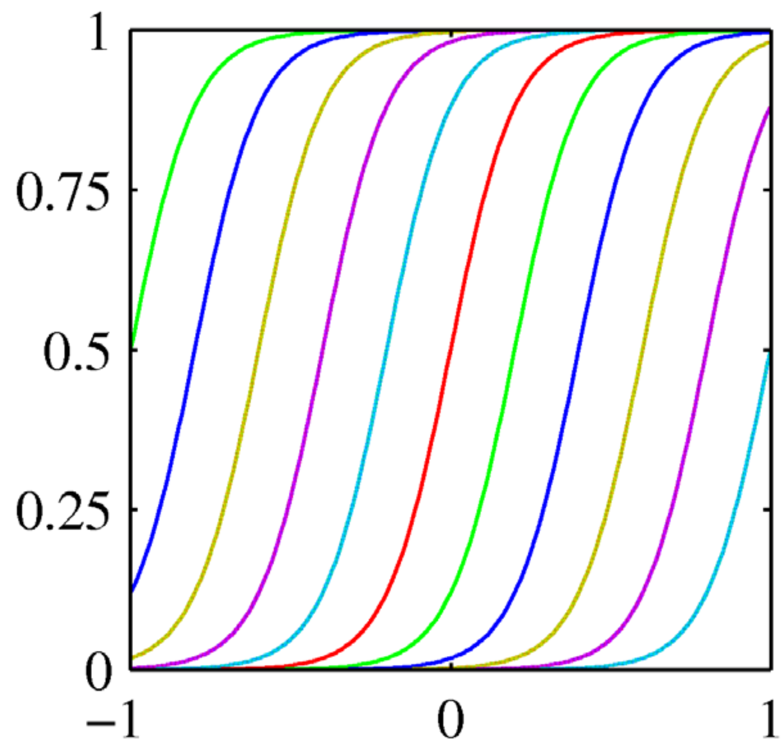
Hàm hồi qui tuyến tính cơ sở (4)

Hàm Sigmoid cơ sở:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

Trong đó:

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$



Cực đại likelihood và bình phương tối thiểu (1)

Giả sử đã có hàm nhiễu Gaussian như sau:

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \quad \text{trong đó} \quad p(\epsilon|\beta) = \mathcal{N}(\epsilon|0, \beta^{-1})$$

Hay có thể viết cách khác:

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}).$$

Cho các quan sát $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ và hàm đích $\mathbf{t} = [t_1, \ldots, t_N]^{\mathrm{T}}$, hàm likelihood:

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}).$$

Cực đại likelihood và bình phương tối thiểu(2)

Lấy ln 2 vế ta có:

$$\begin{aligned}\ln p(\mathbf{t}|\mathbf{w},\beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(t_n|\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n),\beta^{-1}) \\ &= \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) - \beta E_D(\mathbf{w})\end{aligned}$$

Trong đó

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

là hàm tổng bình phương lỗi (sum-of-quares error).

Cực đại likelihood và bình phương tối thiểu(3)

Gradient của log có dạng:

$$\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = \sum_{n=1}^{N} \left\{ t_n - \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) \right\} \phi(\mathbf{x}_n)^{\mathrm{T}} = 0$$

Giải hệ $\nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) = 0$ với biến w ta được:

$$\mathbf{w}_{\mathrm{ML}} = \underbrace{\left(\mathbf{\Phi}^{\mathrm{T}} \mathbf{\Phi} \right)^{-1} \mathbf{\Phi}^{\mathrm{T}}}_{\text{Moore-Penrose pseudo-inverse, } \mathbf{\Phi}^{\dagger}} \mathbf{t}$$

Moore-Penrose pseudo-inverse, $\mathbf{\Phi}^{\dagger}$.

Trong đó:

$$\mathbf{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}.$$

Cực đại likelihood và bình phương tối thiểu (4)

Giả sử y(x,w) =$\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)$ ta có:

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{t_n - w_0 - \sum_{j=1}^{M-1} w_j\phi_j(\mathbf{x}_n)\}^2$$

Cho E_D(w) = 0 ta được:

$$w_0 = \bar{t} - \sum_{j=1}^{M-1} w_j\overline{\phi_j}$$

Với:

$$\bar{t} = \frac{1}{N}\sum_{n=1}^{N} t_n, \qquad \text{và} \qquad \overline{\phi_j} = \frac{1}{N}\sum_{n=1}^{N}\phi_j(\mathbf{x}_n)$$

Cực đại likelihood và bình phương tối thiểu(5)

Từ đó ta đạt được hàm cực đại láng giềng:

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N}\sum_{n=1}^{N}\{t_n - \mathbf{w}_{\mathrm{ML}}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2$$

Bản chất hình học của bình phương tối thiểu

Xét công thức:

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{w}_{\mathrm{ML}} = [\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M]\, \mathbf{w}_{\mathrm{ML}}.$$

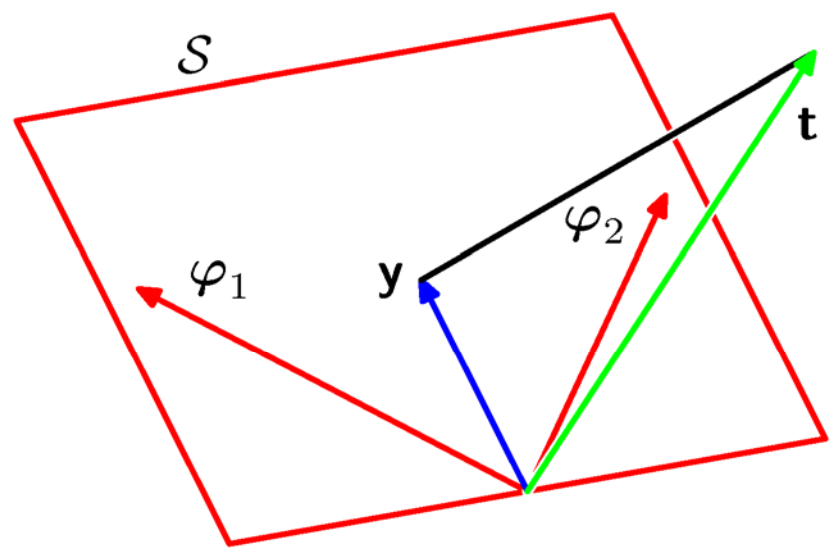
$$\mathbf{y} \in \mathcal{S} \subseteq \mathcal{T} \qquad \mathbf{t} \in \mathcal{T}$$

Trong đó: S là mặt phẳng được xây dựng từ

$\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_M$ (M chiều)

T là không gian N chiều.

w_{ML} là khoảng cách nhỏ nhất từ t với hình chiếu của nó trên S (chính là y).



Sequential Learning (1)

Xử lí theo lô như công thức $\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}$ đòi hỏi phải đưa toàn bộ dữ liệu vào để xử lí cùng lúc $\Rightarrow$ chi phí xử lí lớn (hoặc không đủ bộ nhớ để xử lí). Điều này có thể giải quyết được bằng cách sử dụng các thuật toán tăng cường (sequential hay online)!

$$\begin{aligned}\mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \eta\nabla E_n \\ &= \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n).\end{aligned}$$

Sequential Learning (2)

Có thể sử dụng công thức:

$$\begin{aligned}\mathbf{w}^{(\tau+1)} &= \mathbf{w}^{(\tau)} - \eta \nabla E_n \\ &= \mathbf{w}^{(\tau)} + \eta(t_n - \mathbf{w}^{(\tau)\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}_n))\boldsymbol{\phi}(\mathbf{x}_n).\end{aligned}$$

Trong đó τ là bước lặp thứ τ. $\tau + 1$ biểu thị bước lặp thứ $\tau + 1$.

Cách làm này được gọi là least-mean-squares.

Giá trị η cần được chọn sao cho bảo đảm tính hội tụ của thuật toán!

Regularized Least Squares (1)

Xét hàm lỗi (được trình bày trrong chương 1):

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w})$$

Data term + Regularization term

Tổng bình phương hàm lỗi như sau:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

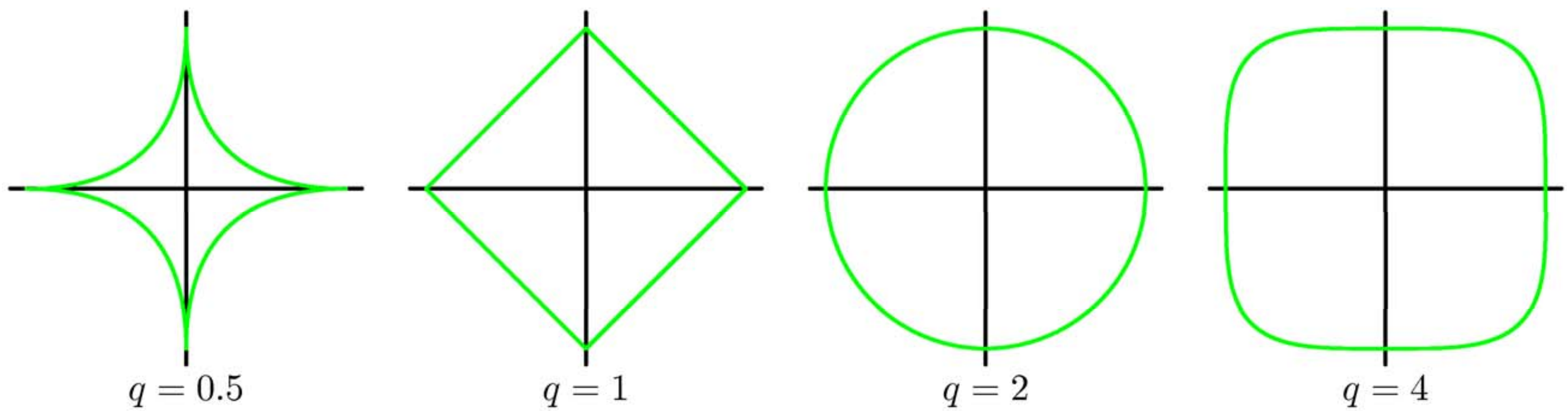
Cực tiểu hóa ta được:

$$\mathbf{w} = \left(\lambda\mathbf{I} + \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}.$$

Regularized Least Squares (2)

Tổng quát hơn, ta có công thức:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$



Lasso

Quadratic

Regularized Least Squares (3)

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

Với q = 2, công thức đã cho trở thành công thức thường dùng (có tên là Quadratic)

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Với q = 1, công thức được gọi là lasso. Trong trường hợp λ đủ lớn, sẽ có một số w_j tiến về 0. Vì vậy, chúng không đóng vai trò gì trong công thức!

Đa đầu ra (1)

Các phần trước xét các trường hợp biến đích t là biến đơn (chỉ chứa 1 thuộc tính). Trong trường hợp T là một ma trận có kích thước MxK, ta có công thức:

$$\begin{aligned} p(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta) &= \mathcal{N}(\mathbf{t}|\mathbf{y}(\mathbf{W}, \mathbf{x}), \beta^{-1}\mathbf{I}) \\ &= \mathcal{N}(\mathbf{t}|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}), \beta^{-1}\mathbf{I}). \end{aligned}$$

Cho quan sát $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ và đích là $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_N]^{\mathrm{T}}$

Ta có hàm log likelihood như sau:

$$\begin{aligned} \ln p(\mathbf{T}|\mathbf{X}, \mathbf{W}, \beta) &= \sum_{n=1}^{N} \ln \mathcal{N}(\mathbf{t}_n|\mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}\mathbf{I}) \\ &= \frac{NK}{2} \ln\left(\frac{\beta}{2\pi}\right) - \frac{\beta}{2} \sum_{n=1}^{N} \left\|\mathbf{t}_n - \mathbf{W}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\right\|^2. \end{aligned}$$

Đa đầu ra (2)

Cực đại hàm trên theo biến W, ta có

$\mathbf{W}_{\mathrm{ML}} = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{T}.$ (giống công thức của 1 target)

Xét 1 target đơn t_k, ta thấy:

$$\mathbf{w}_k = \left(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\right)^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t}_k = \mathbf{\Phi}^{\dagger}\mathbf{t}_k$$

Với $\mathbf{t}_k = [t_{1k}, \ldots, t_{Nk}]^{\mathrm{T}}$, kết quả trên hoàn toàn giống với trường hợp 1 output.