Linear Models for Regression

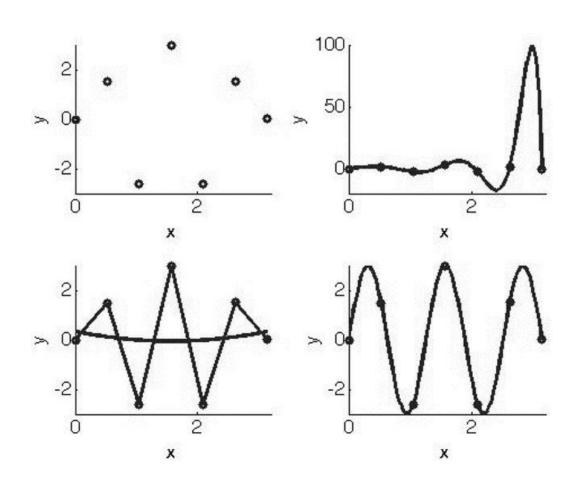
Learning outcomes



- Approximate a function through linear regression with different basis functions
- Compute the pseudo-inverse of a matrix
- Derive the bias-variance decomposition

Back to regression





Linear regression



Unknown function t = f(x)

Samples: $\langle x_i, t_i \rangle$

We approximate the function with a linear model (in the parameters):

$$y(x, w) = w_0 + w_1 x_1 + w_2 x_3 + \dots + w_M x_M$$

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{M} w_i \phi_i(\mathbf{x})$$

Basis functions:

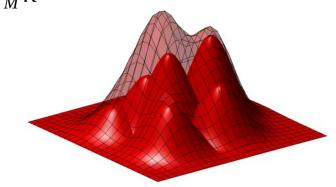
Basis functions



Polynomial basis: $\phi_i(x) = x^i$

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$$

"Gaussian" basis: $\phi_i(x) = \exp\{-\frac{(x-\mu_i)^2}{2s^2}\}$



sigmoid basis:

$$\phi_i(x) = \sigma(\frac{x-\mu}{s}) = \sigma_{\mu,s}$$

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

System of equations



Samples: $\langle x_i, t_i \rangle = \langle x_{i,1}, x_{i,2}, \dots, x_{i,m}, t_i \rangle$

Let's set: $\phi_0 = 1$ So that: $y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{i=1}^{m} \phi_i(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$

For each point: \mathbf{x}_j We denote: $\phi_i(\mathbf{x}_j) = \Phi_{i,j}$

Each point imposes a constraint:

$$\mathbf{w}^{T}\mathbf{\phi}(\mathbf{x}_{i})=t_{i}$$

For N points in the data set, we have N equations.

Can we find a weight vector that satisfies all the constraints?

An Example - data



 $\langle -2, 18.28 \rangle$

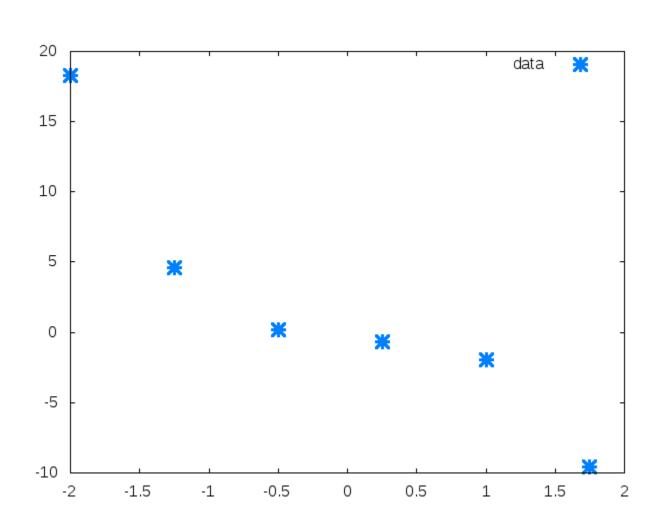
 $\langle -1.25, 4.64 \rangle$

 $\langle -0.5, 0.16 \rangle$

 $\langle 0.25, -0.63 \rangle$

 $\langle 1.75, -9.56 \rangle$

 $\langle 1, -1.95 \rangle$



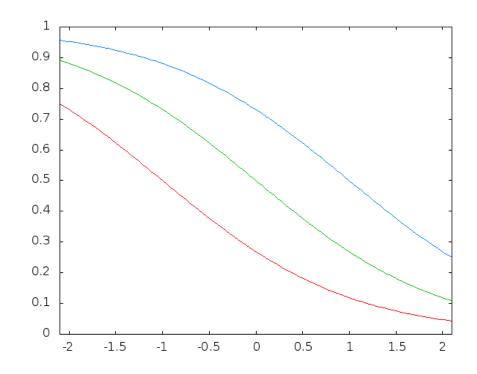
Choose model



How many sigmoids? Centred where? y=?

Let's use these three:

$$y = w_0 + w_1 \sigma_{-1,1}(x) + w_2 \sigma_{0,1}(x) + w_3 \sigma_{1,1}(x)$$



Equations



We create a system of equations by evaluating the basis functions on the data points:

$$w_{0}+w_{1}\sigma_{-1,1}(x)+w_{2}\sigma_{0,1}(x)+w_{3}\sigma_{1,1}(x)=t$$

$$\Phi_{1,3}$$

$$\langle -2,18.28\rangle \qquad w_{0}+w_{1}\sigma_{-1,1}(-2)+w_{2}\sigma_{0,1}(-2)+w_{3}\sigma_{1,1}(-2)=18.28$$

$$\langle -1.25,4.64\rangle \qquad w_{0}+w_{1}\sigma_{-1,1}(-1.25)+w_{2}\sigma_{0,1}(-1.25)+w_{3}\sigma_{1,1}(-1.25)=4.64$$

$$\langle -0.5,0.16\rangle \qquad w_{0}+w_{1}\sigma_{-1,1}(-0.5)+w_{2}\sigma_{0,1}(-0.5)+w_{3}\sigma_{1,1}(-0.5)=0.16$$

$$\langle 0.25,-0.63\rangle \qquad \cdots$$

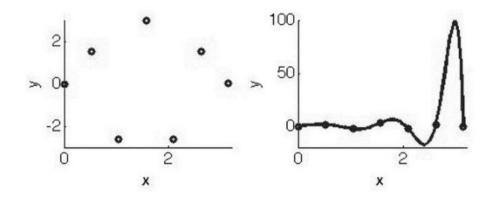
$$\langle 1,-1.95\rangle \qquad \Phi_{3,2} \qquad \cdots$$

In matrix form



$$\begin{bmatrix} \Phi_{1,1} & \Phi_{1,2} & \cdots & \Phi_{1,M+1} \\ \Phi_{2,1} & \Phi_{2,2} & \cdots & \Phi_{2,M+1} \\ \vdots & \vdots & \vdots & \vdots \\ \Phi_{N,1} & \Phi_{N,2} & \cdots & \Phi_{N,M+1} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$$
 or: $\mathbf{\Phi} \mathbf{w} = \mathbf{t}$

If Φ square and nonsingular: $w = \Phi^{-1}t$



No error, the fitted function goes through every point

Exact solution



If we only had 4 points, we could find an exact solution, that is a function that goes through the points:

$$\langle -2, 18.28 \rangle$$

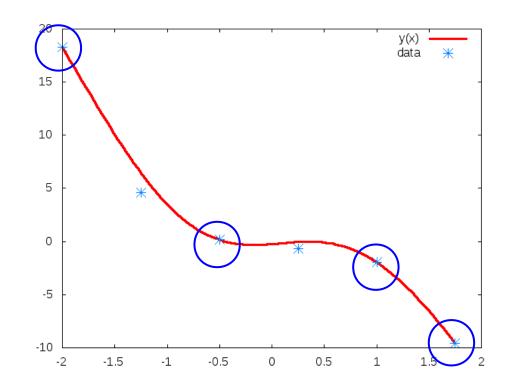
$$\langle -0.5, 0.16 \rangle$$

$$\langle 1, -1.95 \rangle$$

$$\langle 1.75, -9.56 \rangle$$

$$w = \Phi^{-1} t$$

$$\vec{w} = \langle -34.99, 184.249, -312.42, 209.20 \rangle$$



Overdetermined vector



If vector w overdetermined (more equations than variables) \rightarrow no exact solution available.

Then (guess what), define an error and minimize.

You would like:

$$\boldsymbol{\phi}_{i}^{T}(\boldsymbol{x}_{i})\boldsymbol{w}=\boldsymbol{t}_{i} \ \forall \boldsymbol{x}_{i}$$

So a good error seems to be the norm of the difference between the left and right-end side of that equation:

$$E = \frac{1}{2} \sum_{i}^{N} ||\boldsymbol{\phi}_{i}^{T} \boldsymbol{w} - \boldsymbol{t}_{i}||^{2}$$

Sum-of-squares error



$$E = \frac{1}{2} \sum_{i=1}^{N} (y_i - t_i)^2 = \frac{1}{2} \sum_{i=1}^{N} (\phi_i^T w - t_i)^2$$

As usual, compute the gradient:

$$\nabla E = ?$$

Sum-of-squares error



$$E = \frac{1}{2} \sum_{i=1}^{N} (y_i - t_i)^2 = \frac{1}{2} \sum_{i=1}^{N} (\phi_i^T w - t_i)^2$$

As usual, compute the gradient:

$$\nabla E = \sum_{i=1}^{N} (\boldsymbol{\phi}_{i}^{T} \boldsymbol{w} - \boldsymbol{t}_{i}) \boldsymbol{\phi}_{i} = \boldsymbol{\Phi}^{T} (\boldsymbol{\Phi} \boldsymbol{w} - \boldsymbol{t})$$

Least squares solution



$$\Phi^{T}(\Phi\vec{w}-\vec{t})=0 \Rightarrow \Phi^{T}\Phi w=\Phi^{T}t$$

Necessary condition for a vector w to be a minimum

 $\Phi^T \Phi$ Is a square matrix.

If it is also non singular, we can invert it and solve the equation above:

$$\mathbf{w} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

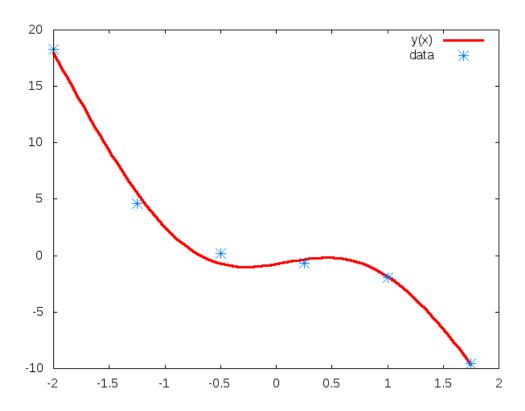
where $\Phi_p = (\Phi^T \Phi)^{-1} \Phi^T$ is the *pseudoinverse* of Φ

Least-squares solution



$$\mathbf{w} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{t}$$

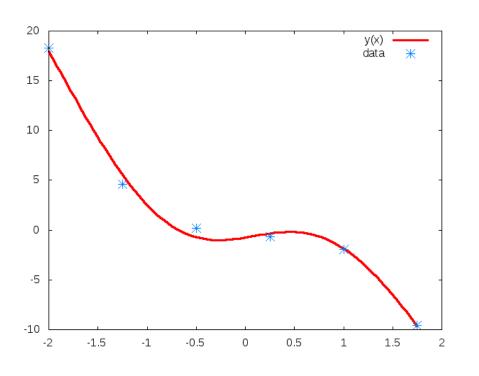
$$w = \langle -36.53, 198.30, -339.45, 225.09 \rangle$$



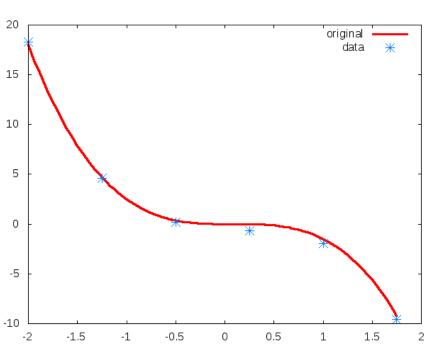
Comparison



Least-squares



Original function



$$f(x)=0.5x^2-2x^3$$

Given the dataset: $\{(0,0), (0,1), (1,0)\}$, find the least-squares solution for the parameters in the regression of the function: $y=w_1+w_2x^2$

- 1) Evaluate the bases on the points: $\Phi = ?$
- 2) Compute: $\Phi^T \Phi = ?$
- 3) Invert: $(\Phi^T \Phi)^{-1} = ?$
- 4) Compute the pseudo-inverse: $\Phi_p = (\Phi^T \Phi)^{-1} \Phi^T = ?$
- 5) Compute w! $w = \Phi_p t = ?$



Given the dataset: $\{(0,0), (0,1), (1,0)\}$, find the least-squares solution for the parameters in the regression of the function: $y=w_1+w_2x^2$

1) Evaluate the bases on the points:
$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$



Given the dataset: {(0,0), (0,1), (1,0)}, find the least-squares solution for the parameters in the regression of the function: $y = w_1 + w_2 x^2$

1) Evaluate the bases on the points:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

2) Compute: $\Phi^T \Phi = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

Given the dataset: $\{(0,0), (0,1), (1,0)\}$, find the least-squares solution for the parameters in the regression of the function: $y = w_1 + w_2 x^2$

1) Evaluate the bases on the points:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

2) Compute:
$$\mathbf{\Phi}^T \mathbf{\Phi} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

3) Invert:
$$(\mathbf{\Phi}^T \mathbf{\Phi})^{-1} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$



Given the dataset: $\{(0,0), (0,1), (1,0)\}$, find the least-squares solution for the parameters in the regression of the function: $y = w_1 + w_2 x^2$

1) Evaluate the bases on the points:

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

2) Compute:
$$\Phi^T \Phi = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

3) Invert:
$$(\mathbf{\Phi}^T \mathbf{\Phi})^{-1} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$$

4) Compute the pseudo-inverse:
$$\Phi_p = (\Phi^T \Phi)^{-1} \Phi^T = \begin{vmatrix} 0.5 & 0.5 & 0 \\ -0.5 & -0.5 & 1 \end{vmatrix}$$



Given the dataset: $\{(0,0), (0,1), (1,0)\}$, find the least-squares solution for the parameters in the regression of the function: $y = w_1 + w_2 x^2$

- $\mathbf{\Phi} = \begin{vmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{vmatrix}$ 1) Evaluate the bases on the points:
- 2) Compute: $\Phi^T \Phi = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{vmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \end{vmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$
- 3) Invert: $(\mathbf{\Phi}^T \mathbf{\Phi})^{-1} = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 1.5 \end{bmatrix}$
- 4) Compute the pseudo-inverse: $\Phi_p = (\Phi^T \Phi)^{-1} \Phi^T = \begin{bmatrix} 0.5 & 0.5 & 0 \\ -0.5 & -0.5 & 1 \end{bmatrix}$ 5) Compute w! $\mathbf{w} = \Phi_p \mathbf{t} = \begin{bmatrix} 0.5 & 0.5 & 0 \\ -0.5 & -0.5 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}$

Sequential learning



The closed-form solution requires to process the whole data set.

If
$$E = \sum_{n} E_n$$

As with neural networks, it is also possible to consider one point at a time

$$w^{(t+1)} = w^{(t)} - \eta \nabla E_n$$

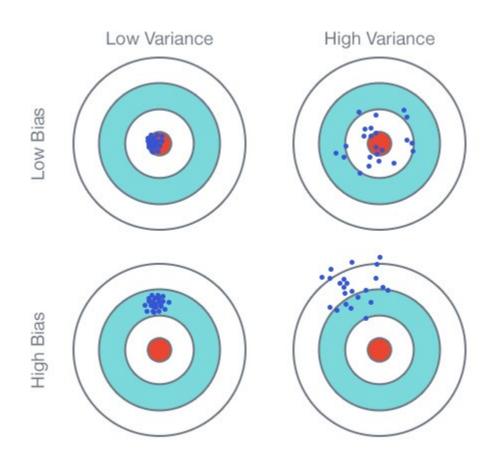
For a least-squares problem:

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \boldsymbol{\eta} \, \boldsymbol{\phi}_n^T (\boldsymbol{\phi}_n^T \, \boldsymbol{w}^{(t)} - \boldsymbol{t}_n)$$

known as the Least-mean-squares (LMS) algorithm

Bias and Variance





Bias-Variance decomposition



Underlying function plus noise: $f(\vec{x}) + \epsilon$

Generates different datasets D

Simplified notation: $\hat{y} \equiv y(\vec{x}; D)$ $f \equiv f(\vec{x})$

$$\hat{y} \equiv y(\vec{x}; D)$$

$$f \equiv f(\vec{x})$$

Expected value of the error with respect to datasets D (I excluded noise from these calculations, the book has a version with noise):

Bias-Variance decomposition



$$=E_{D}[(\hat{y}-E_{D}[\hat{y}])^{2}]+2E_{D}[(\hat{y}-E_{D}[\hat{y}])(E_{D}[\hat{y}]-f)]+E_{D}[(E_{D}[\hat{y}]-f)^{2}]$$

$$=E_{D}[(\hat{y}-E_{D}[\hat{y}])^{2}]+E_{D}[(E_{D}[\hat{y}]-f)^{2}]$$

$$\text{Variance}$$

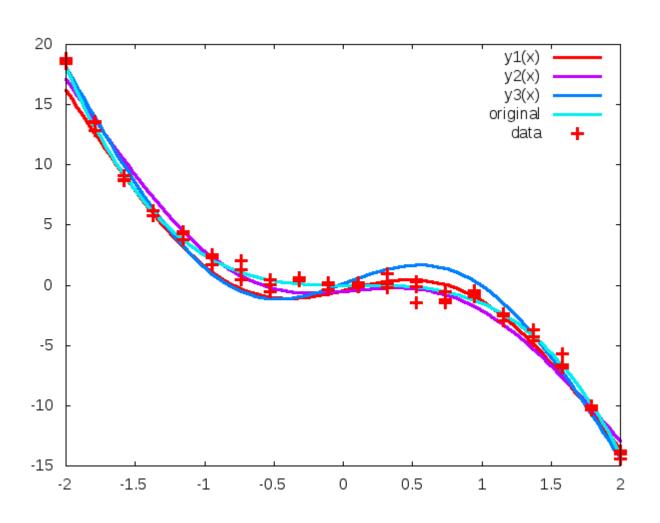
$$\text{Bias^2}$$

Recall that the variance of a random variable is: $var(X) = E[(X - E[X])^2]$

Not limited to regression with mean squared error: general phenomenon

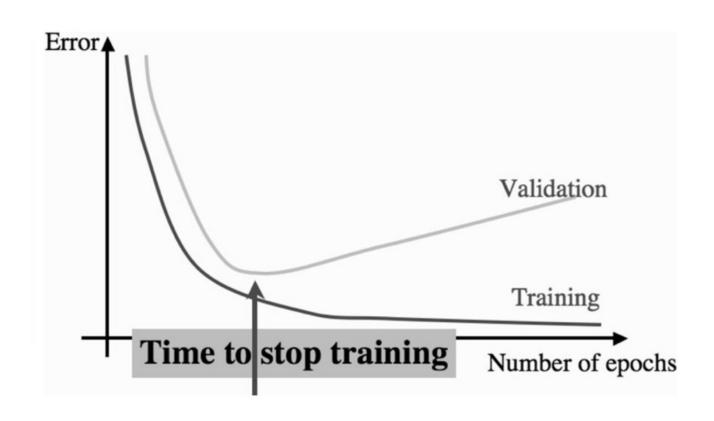
Bias-Variance

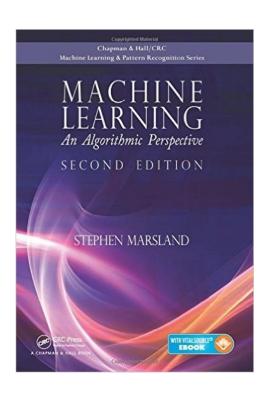




On training and validation error







Chapter 2.5 and 3.5