

Model Linear untuk Regresi

Dr. rer. nat. Hendri Murfi

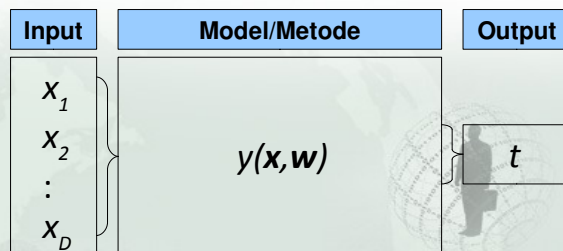
Intelligent Data Analysis (IDA) Group

Departemen Matematika, Universitas Indonesia – Depok 16424

Telp. +62-21-7862719/7863439, Fax. +62-21-7863439, Email. hendri@ui.ac.id

Machine Learning

Tahapan Umum Proses



Diberikan data pelatihan (*training data*), yaitu $\mathbf{x}_i$ dan/atau $\mathbf{t}_i$, $i = 1$ sd N

- *Preprocessing*: pemilihan/ekstraksi fitur dari data, misal $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iD})^T$
- *Learning*: penentuan parameter metode, misal $\mathbf{w}$, berdasarkan data pelatihan
- *Testing*: pengujian metode dengan data baru. Data penguji (*testing data*) tersebut harus dilakukan *preprocessing* yang sama dengan data pembelajaran sebelum dieksekusi oleh metode

Learning

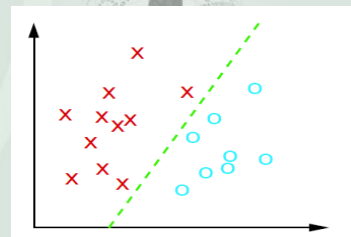
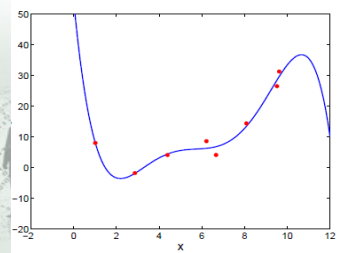
Diberikan data pelatihan $\mathbf{x}_i$, $i = 1$ sd N , dan/atau $\mathbf{t}_i$, $i = 1$ as N

- **Supervised Learning.** Data pelatihan disertai target, yaitu $\{\mathbf{x}_i, \mathbf{t}_i\}$, $i = 1$ sd N . Tujuan pembelajaran adalah membangun model yang dapat menghasilkan output yang benar untuk suatu data input, misal untuk regresi, klasifikasi, regresi ordinal, ranking, dll
- **Unsupervised Learning.** Data pelatihan tidak disertai target, yaitu $\mathbf{x}_i$, $i = 1$ sd N . Tujuan pembelajaran adalah membangun model yang dapat menemukan komponen/variabel/fitur tersembunyi pada data pelatihan, yang dapat digunakan untuk: pengelompokan (*clustering*), reduksi dimensi (*dimension reduction*), rekomendasi, dll

3

Supervised Learning

- **Regresi**
 - Nilai output $\mathbf{t}_i$ bernilai kontinu (riil)
 - Bertujuan memprediksi output dengan akurat untuk data baru
- **Klasifikasi**
 - Nilai output $\mathbf{t}_i$ bernilai diskrit (kelas)
 - Bertujuan mengklasifikasi data baru dengan akurat



4

Regresi

Model Linear

- Model linear adalah kombinasi linear dari fungsi nonlinear dari variabel input (fungsi basis):

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

dimana $\mathbf{x} = (x_1, x_2, \dots, x_D)^T$ adalah variabel input, dan $\mathbf{w} = (w_0, w_1, \dots, w_{M-1})^T$ adalah parameter, $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \dots, \phi_{M-1}(\mathbf{x}))^T$ adalah vektor *fungsi basis* $\phi_j(\mathbf{x})$, M adalah jumlah total parameter dari model

- Biasanya, $\phi_0(\mathbf{x}) = 1$, sehingga w_0 berfungsi sebagai bias
- Ada banyak pilihan yang mungkin untuk fungsi basis $\phi(\mathbf{x})$, misal fungsi linear, fungsi polinomial, fungsi gaussian, fungsi sigmoidal, dll

5

Regresi Linear Sederhana

Fungsi Basis Polinomial

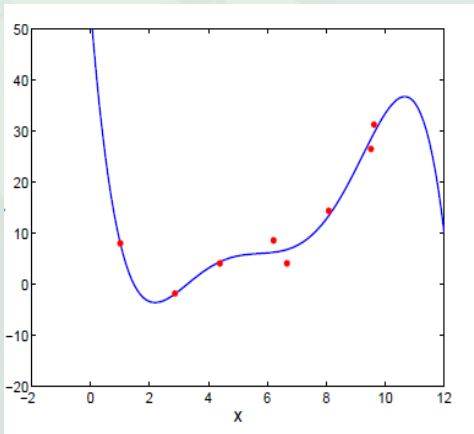
- Regresi linear sederhana (*simple linear regression*) adalah masalah regresi dengan variabel input x berdimensi satu. Misal kita menggunakan polinomial $\phi_j(x) = x^j$ sebagai fungsi basis, dan $M = M-1$, maka bentuk umum dari regresi linear sederhana tersebut adalah:

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

6

Regresi Linear Sederhana

Polynomial Curve Fitting



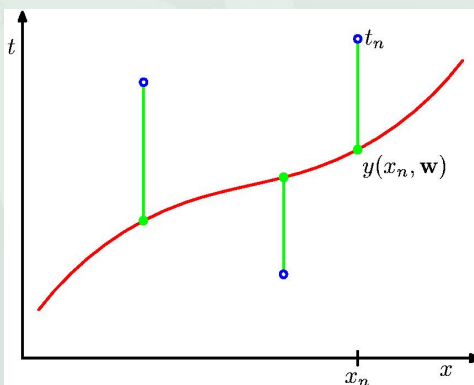
Diberikan data pelatihan $\{x_i, t_i\}, i = 1 \text{ sd } N$

- Masalah: bagaimana mendapatkan kurva polinomial yang cocok untuk data pelatihan tersebut
- Solusi: mencari kurva polinomial yang memiliki kesalahan (*error*) terkecil pada data pelatihan tersebut
- Persoalan ini sering juga disebut sebagai *polynomial curve fitting*

7

Regresi Linear Sederhana

Fungsi Error



- Salah satu fungsi error yang sering digunakan adalah fungsi *sum-of-squares error* sbb:

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

- Salah satu metode yang digunakan untuk mencari nilai $\mathbf{w}$ yang meminimumkan fungsi error adalah metode kuadrat terkecil (*least squares*)

8

Regresi Linear Sederhana

Metode Kuadrat Terkecil

- Setelah penurunan $E(\mathbf{w})$ terhadap $\mathbf{w}$, maka persoalan penentuan nilai parameter $\mathbf{w}$ menjadi persoalan penentuan solusi sistem persamaan linear:

$$A\mathbf{w} = \mathbf{t}$$

dimana

$$A = \begin{bmatrix} \sum_{n=1}^N 1 & \sum_{n=1}^N x_n & \cdots & \sum_{n=1}^N x_n^M \\ \sum_{n=1}^N x_n & \sum_{n=1}^N x_n^2 & \cdots & \sum_{n=1}^N x_n^{M+1} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{n=1}^N x_n^M & \sum_{n=1}^N x_n^{M+1} & \cdots & \sum_{n=1}^N x_n^{2M} \end{bmatrix}$$

$$\mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_M \end{bmatrix}$$

$$\mathbf{t} = \begin{bmatrix} \sum_{n=1}^N t_n \\ \sum_{n=1}^N x_n t_n \\ \vdots \\ \sum_{n=1}^N x_n^M t_n \end{bmatrix}$$

9

Regresi Linear Sederhana

Contoh Kasus

Seorang ahli biologi telah melakukan eksperimen sebanyak 7 kali untuk melihat pertumbuhan bakteri berdasarkan kadar Nitrogen, dan diperoleh kondisi sbb:

Kadar Nitrogen (gram)	3	4	6	7	8	9
Pertumbuhan Bakteri	1	3	4	6	8	8

Tentukan regresi linear polinomial berorde 1 berdasarkan data tsb. Selanjutnya, prediksi pertumbuhan bakteri jika diberikan Nitrogen sebanyak 5 gram.

Solusi:

Dari persoalan diatas diketahui x = kadar nitrogen, t = pertumbuhan bakteri, $N=6$ dan $M=1$, sehingga:

$$A = \begin{bmatrix} 6 & 37 \\ 37 & 255 \end{bmatrix}, \quad \mathbf{t} = \begin{bmatrix} 30 \\ 217 \end{bmatrix}, \quad \text{dan } \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \text{ adalah solusi SPL } A\mathbf{w} = \mathbf{t}, \text{ yaitu } \mathbf{w} = \begin{bmatrix} -2.35 \\ 1.19 \end{bmatrix}$$

dan model linear yang dihasilkan adalah $y(x) = -2.35 + 1.19x$. Sementara prediksi pertumbuhan bakteri untuk 5 gram Nitrogen adalah $y(5) = -2.35 + 1.19 \cdot 5 = 3.6$

10

Regresi Linear Sederhana

Contoh Kasus: Menggunakan Weka

Format data*:

@RELATION bakteri

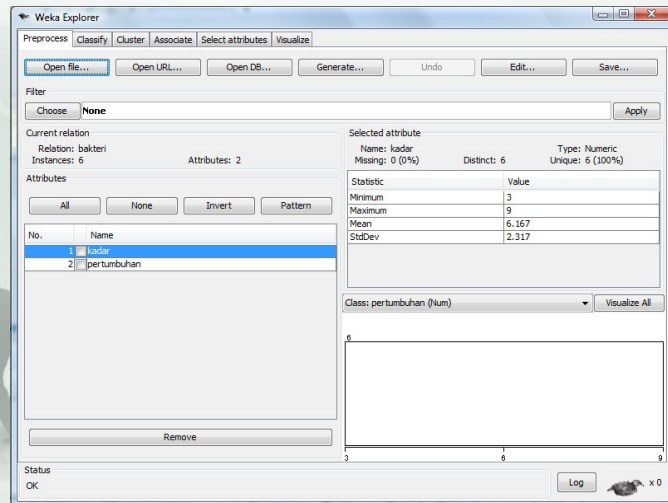
@ATTRIBUTE kadar NUMERIC

@ATTRIBUTE pertumbuhan NUMERIC

@DATA

3, 1
4, 3
6, 4
7, 6
8, 8
9, 8

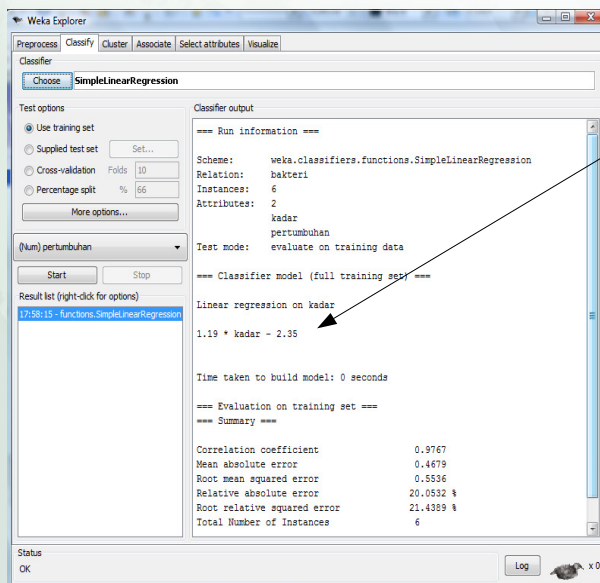
*Disimpan dalam file dengan ekstensi arff (misal: bakteri.arff)



11

Regresi Linear Sederhana

Contoh Kasus: Menggunakan Weka



Model hasil:

$$y(x) = -2.35 + 1.19x$$

12

Pemilihan Model

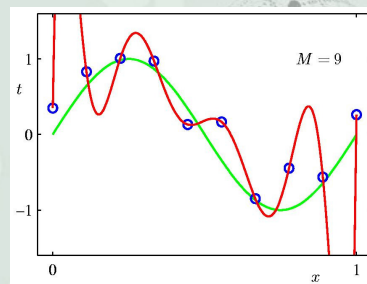
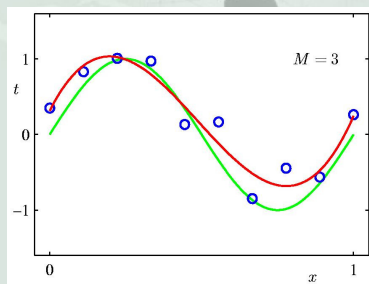
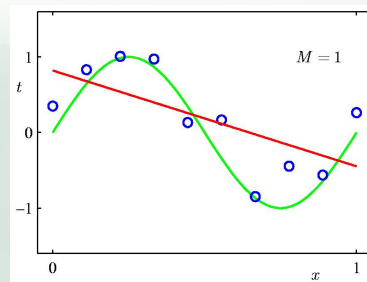
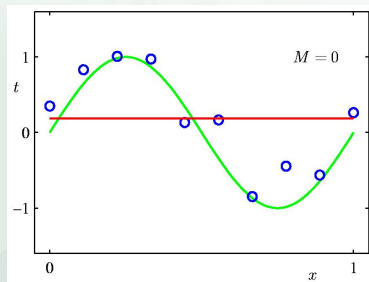
- Karakteristik model regresi linear polinomial ditentukan oleh nilai M (orde polinomial atau jumlah parameter). Pemilihan nilai M yang optimal dikenal juga dengan istilah pemilihan model (*model selection*)

$$y(x, \mathbf{w}) = w_0 + w_1x + w_2x^2 + \dots + w_Mx^M = \sum_{j=0}^M w_jx^j$$

13

Pemilihan Model

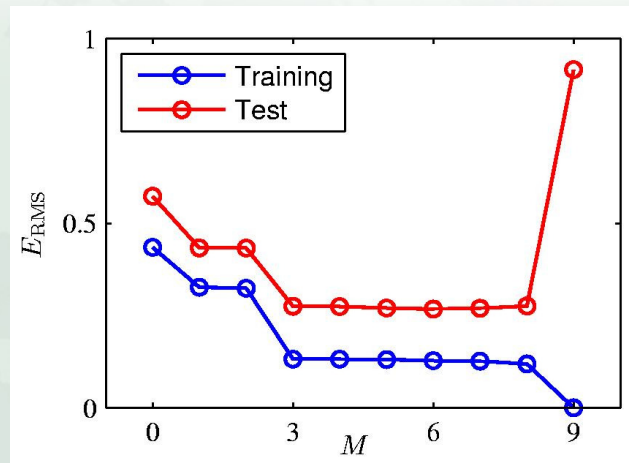
Under-fitting dan Over-fitting



14

Pemilihan Model

Under-fitting dan Over-fitting

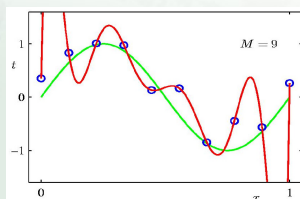


Root-Mean-Square (RMS) Error:

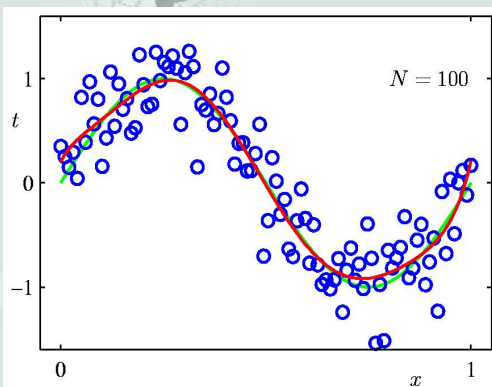
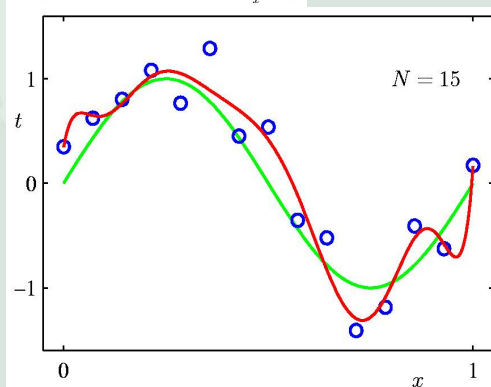
$$E_{\text{RMS}} = \sqrt{2E(\mathbf{w}^*)/N}$$

15

Parameter vs Data



- Jumlah data pembelajaran seharusnya tidak lebih sedikit dari jumlah parameter



16

Regularisasi

- Pada aplikasi praktis, kita sering menemukan kondisi dimana untuk persoalan yang kompleks ketersediaan data pembelajaran terbatas.
- Salah satu teknik yang digunakan untuk mengontrol fenomena *over-fitting* adalah regularisasi (*regularization*), yaitu dengan cara menambah finalti ke fungsi error.

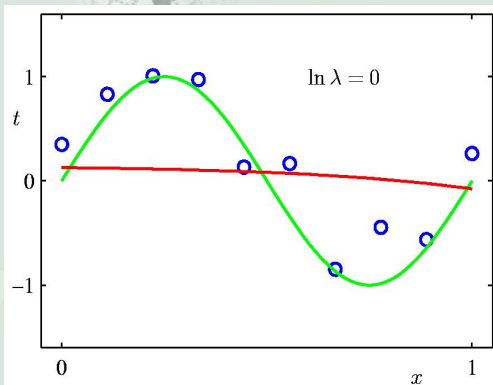
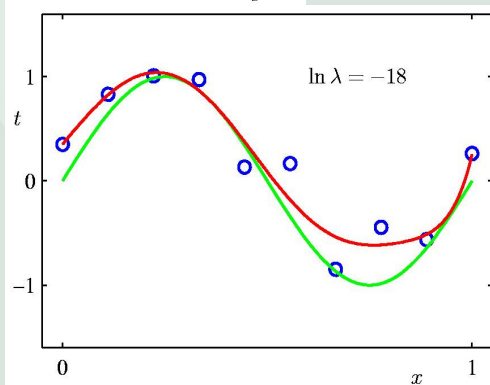
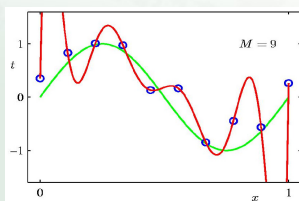
	$M=0$	$M=1$	$M=3$	$M=9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

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

17

Regularisasi

Penghalusan Kurva



18

Regularisasi

Pengecilan Nilai Bobot

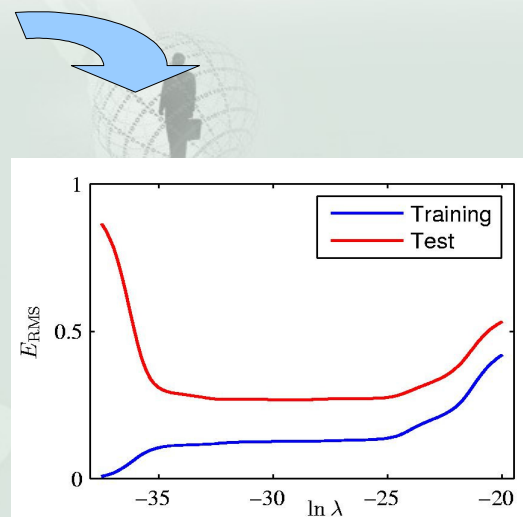
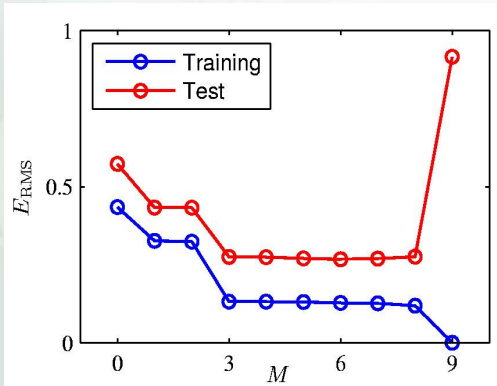
	$M = 0$	$M = 1$	$M = 3$	$M = 9$
w_0^*	0.19	0.82	0.31	0.35
w_1^*		-1.27	7.99	232.37
w_2^*			-25.43	-5321.83
w_3^*			17.37	48568.31
w_4^*				-231639.30
w_5^*				640042.26
w_6^*				-1061800.52
w_7^*				1042400.18
w_8^*				-557682.99
w_9^*				125201.43

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

19

Regularisasi

Mengatasi over-fitting



20

Regresi Linear Umum

- Fungsi *sum square error* adalah

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

nilai bobot $\mathbf{w}$ yang meminimum fungsi error adalah

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

dimana

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

21

Regresi Linear Umum

Regularisasi

- Fungsi *regularized sum square error* adalah

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

nilai bobot $\mathbf{w}$ yang meminimum fungsi error adalah

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

dimana

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

22

Referensi

- Bishop, C. H., *Pattern Recognition and Machine Learning*, Springer, 2006 (Bab 1.1, Bab 1.3, Bab 1.4, Bab 3.1)