Regresi

Ikhtisar

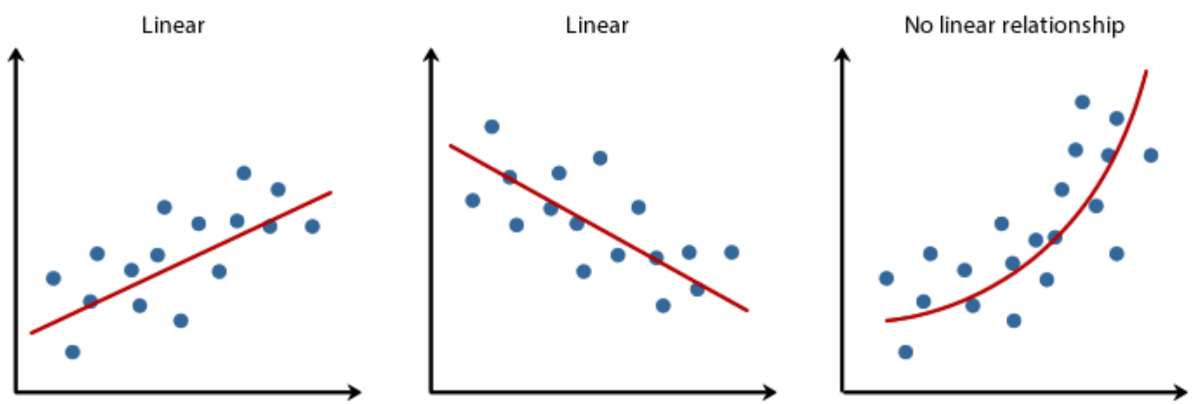
- Pengantar
- Regresi Linear
- Regresi Non-linear (Polinomial)
- Regresi dengan Regularisasi
- Regresi menggunakan ANN

Pengantar

Pengantar

- **Pengertian**: teknik pembelajaran yang fokus pada pada hubungan antara variabel independen (prediktor) dan variabel dependen (output), yang mana variabel dependennya berupa nilai **kontinyu**.

- **Contoh**:
  - ✓ Memprediksi harga sebuah produk berdasarkan atribut-atribut yang diberikan
  - ✓ Memprediksi *bounding box* pada deteksi objek
  - ✓ Memprediksi keterlambatan sebuah pesawat



Overview regresi

- Membutuhkan label
- Outputnya **kontinyu**
- Contoh kasus:
  - ✓ Memprediksi harga rumah, dengan diberikan **fitur** seperti: luas rumah, tingkat kriminalitas, akses ke transportasi umur, dsb.
  - ✓ Contoh data: data '*housing*' dari UC Irvine

0.00632	18.00	2.310	0	0.5380	6.5750	65.20	4.0900	1	296.0	15.30	396.90	4.98	24.00
0.02731	0.00	7.070	0	0.4690	6.4210	78.90	4.9671	2	242.0	17.80	396.90	9.14	21.60
0.02729	0.00	7.070	0	0.4690	7.1850	61.10	4.9671	2	242.0	17.80	392.83	4.03	34.70
0.03237	0.00	2.180	0	0.4580	6.9980	45.80	6.0622	3	222.0	18.70	394.63	2.94	33.40
0.06905	0.00	2.180	0	0.4580	7.1470	54.20	6.0622	3	222.0	18.70	396.90	5.33	36.20
0.02985	0.00	2.180	0	0.4580	6.4300	58.70	6.0622	3	222.0	18.70	394.12	5.21	28.70
0.08829	12.50	7.870	0	0.5240	6.0120	66.60	5.5605	5	311.0	15.20	395.60	12.43	22.90

Fitur/atribut

Label/ground truth = harga rumah (unit = $1k)

Regresi Linear

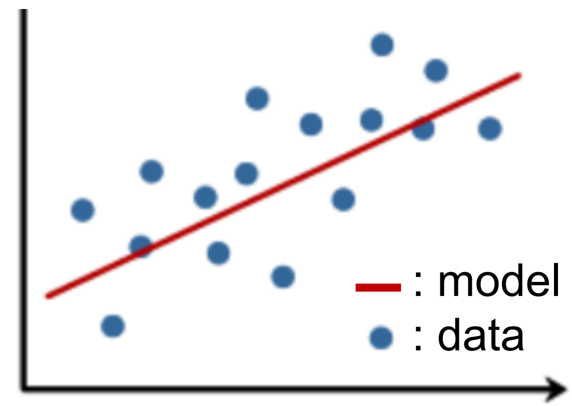
Regresi Linear

- Regresi linear dapat kita modelkan

$$y(\boldsymbol{x}) = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

di mana:

a_i = koefisien regresi
x_i = atribut / fitur / var. independen
y = output prediksi / var. dependen



Dalam notasi matriks, dapat kita tuliskan:

$$y(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{x}, \text{ dengan } \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \text{ dan } \boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Untuk pasangan input-ouput sebanyak m data, dapat kita tuliskan:

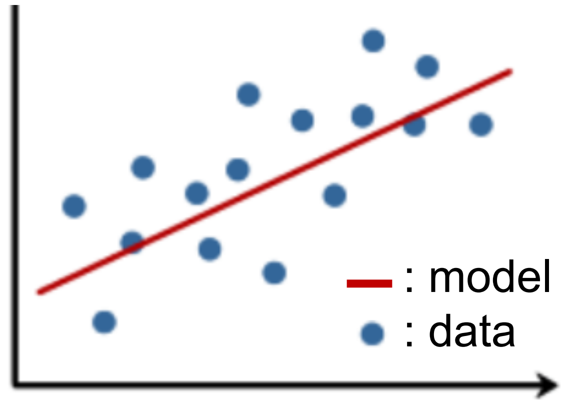
$$\begin{bmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_m) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \Leftrightarrow y(\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{a}$$

Disebut sebagai matriks desain (*design matrix*)

Regresi Linear

- Untuk melatih model regresi, dapat kita lakukan dengan **meminimalkan** *loss function:*

$$f(\boldsymbol{a}) = \sum_{i=1}^{m} (y(x_i) - \bar{y}_i)^2$$



- Atau dalam notasi matriksnya:

$$f(\boldsymbol{a}) = (\boldsymbol{y}(\boldsymbol{x}) - \bar{\boldsymbol{y}})^2 = (\boldsymbol{X}\boldsymbol{a} - \bar{\boldsymbol{y}})^2,$$

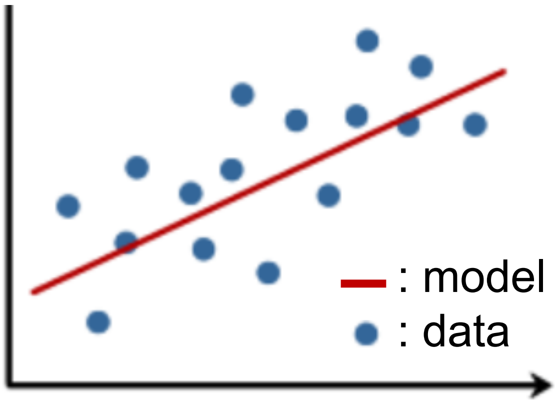
Ingat → $y(\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{a} \Leftrightarrow \begin{bmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_m) \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix}$

- Untuk mencari $\boldsymbol{a}$ yang meminimalkan *loss function* $f(\boldsymbol{a})$, dapat dilakukan dengan $\frac{d(f(\boldsymbol{a}))}{d\boldsymbol{a}} = 0$

Regresi Linear

- Untuk melatih model regresi, dapat kita lakukan dengan **meminimalkan** *loss function:*

$$f(\boldsymbol{a}) = \sum_{i=1}^{m} (y(x_i) - \bar{y}_i)^2$$



- Atau dalam notasi matriksnya:

$$f(\boldsymbol{a}) = (y(\boldsymbol{x}) - \bar{y}_i)^2 = (\boldsymbol{Xa} - \bar{\boldsymbol{y}})^2 = (\boldsymbol{Xa})^2 - 2(\boldsymbol{Xa})^T\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^2$$

- Untuk mencari $\boldsymbol{a}$ yang meminimalkan *loss function* $f(\boldsymbol{a})$, dapat dilakukan dengan $\frac{d(f(\boldsymbol{a}))}{d\boldsymbol{a}} = 0$

$$\frac{d(f(\boldsymbol{a}))}{d\boldsymbol{a}} = \frac{d((\boldsymbol{Xa})^2 - 2(\boldsymbol{Xa})^T\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^2)}{\boldsymbol{a}} = 0$$

$$0 = 2\boldsymbol{X}^T\boldsymbol{Xa} - 2\boldsymbol{X}^T\bar{\boldsymbol{y}}$$

$$\boldsymbol{X}^T\boldsymbol{Xa} = \boldsymbol{X}^T\bar{\boldsymbol{y}} \iff \boxed{\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\bar{\boldsymbol{y}}}$$

Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}; \bar{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

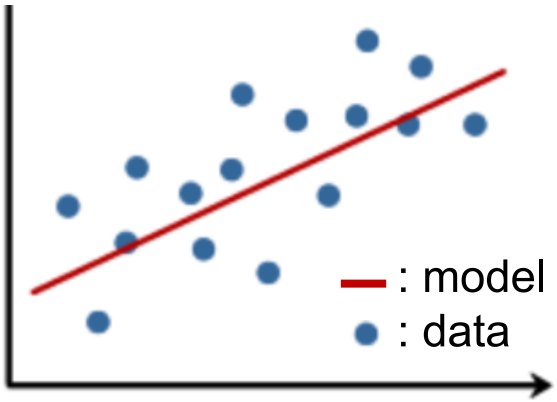
Regresi Linear

- Kita dapatkan koefisien regresi:

$$\boldsymbol{a} = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\overline{\boldsymbol{y}}$$

- Maka prediksi regresinya adalah:

$$y(\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{a}$$



Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}; \overline{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

Regresi Linear

- Diberikan data, cari model regresinya!

Observasi	Temperatur (x_1, $satuan = C$)	Kec. katalis (x_2, $satuan = kg/jam$)	Viskositas
1	80	8	2256
2	93	9	2340
3	100	10	2426
4	82	12	2293
5	90	11	2330
6	99	8	2368
7	81	8	2368
8	96	10	2250
9	94	12	2409
10	93	11	2364
11	97	13	2440
12	95	11	2364
13	100	8	2404
14	85	12	2317
15	86	9	2309
16	87	12	2328

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 1 & 80 & 8 \\ 1 & 93 & 9 \\ \vdots & \vdots & \vdots \\ 1 & 87 & 12 \end{bmatrix}$$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 16 & 1458 & 164 \\ 1458 & 133560 & 14946 \\ 164 & 14946 & 1726 \end{bmatrix}$$

$$\boldsymbol{X}^T\overline{\boldsymbol{y}} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 80 & 93 & \cdots & 87 \\ 8 & 9 & \cdots & 12 \end{bmatrix} \begin{bmatrix} 2256 \\ 2340 \\ \vdots \\ 2328 \end{bmatrix}$$

$$\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}} = \begin{bmatrix} 1566{,}08 \\ 7{,}62 \\ 8{,}58 \end{bmatrix}$$

- Jadi model regresinya:

$$y = 1566{,}08 + 7{,}62x_1 + 8{,}58x_2$$

Regresi Non-linear

Regresi Polinomial (Non-linear)

- Regresi non-linear dapat kita modelkan

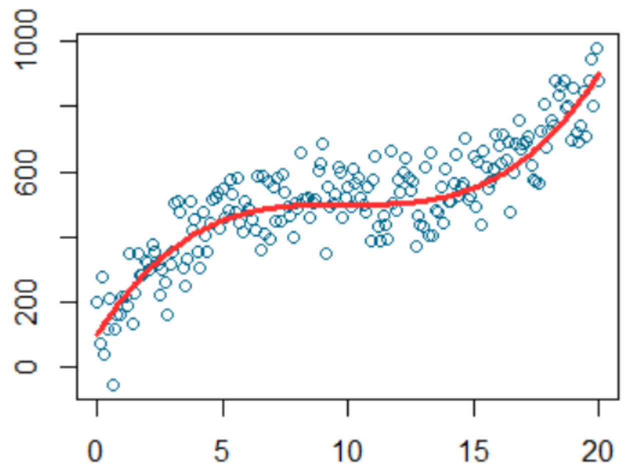
$$y(\boldsymbol{x}) = a_0 + a_1 x_1 + a_2 x_1^2 + \cdots + a_n x_1^n$$

di mana:

a_i = koefisien regresi

x_i = atribut / fitur / var. independen

y = output prediksi / var. dependen



Dalam notasi matriks, dapat kita tuliskan:

$$y(\boldsymbol{x}) = \boldsymbol{a}^T \boldsymbol{x}, \text{ dengan } \boldsymbol{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} \text{ dan } \boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_n^n \end{bmatrix}$$

Untuk pasangan input-ouput sebanyak m data, dapat kita tuliskan:

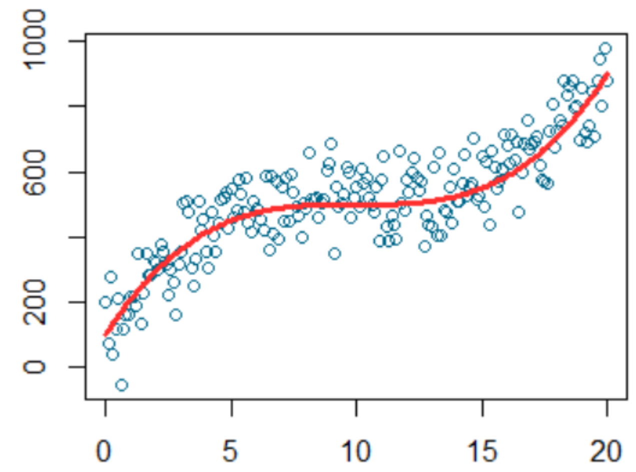
$$\begin{bmatrix} y(x_1) \\ y(x_2) \\ \vdots \\ y(x_m) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{11}^2 & \dots & x_{11}^n \\ 1 & x_{21} & x_{21}^2 & \dots & x_{21}^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m1}^2 & \dots & x_{m1}^n \end{bmatrix}} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \Leftrightarrow y(\boldsymbol{x}) = \boldsymbol{X}\boldsymbol{a}$$

Disebut sebagai matriks desain (*design matrix*)

Regresi Polinomial (Non-linear)

- Untuk melatih model regresi, dapat kita lakukan dengan **meminimalkan** *loss function:*

$$f(\boldsymbol{a}) = \sum_{i=1}^{m} (y(x_i) - \bar{y}_i)^2$$



- Atau dalam notasi matriksnya:

$$f(\boldsymbol{a}) = (y(\boldsymbol{x}) - \bar{y}_i)^2 = (\boldsymbol{Xa} - \bar{\boldsymbol{y}})^2 = (\boldsymbol{Xa})^2 - 2(\boldsymbol{Xa})^T\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^2$$

- Untuk mencari $\boldsymbol{a}$ yang meminimalkan *loss function* $f(\boldsymbol{a})$, dapat dilakukan dengan $\frac{d(f(\boldsymbol{a}))}{\boldsymbol{a}} = 0$

$$\frac{d(f(\boldsymbol{a}))}{\boldsymbol{a}} = \frac{d((\boldsymbol{Xa})^2 - 2(\boldsymbol{Xa})^T\bar{\boldsymbol{y}} + \bar{\boldsymbol{y}}^2)}{\boldsymbol{a}} = 0$$

$$0 = 2\boldsymbol{X}^T\boldsymbol{Xa} - 2\boldsymbol{X}^T\bar{\boldsymbol{y}}$$

$$\boldsymbol{X}^T\boldsymbol{Xa} = \boldsymbol{X}^T\bar{\boldsymbol{y}} \iff \boxed{\boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\bar{\boldsymbol{y}}}$$

Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{11}^2 & \dots & x_{11}^n \\ 1 & x_{21} & x_{21}^2 & \dots & x_{21}^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m1}^2 & \dots & x_{m1}^n \end{bmatrix}; \bar{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

Perbandingan Regresi Linear dan Non-linear

- Formula koefisien regresi linear:

$$\boxed{\boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}}}$$

Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}; \overline{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

- Formula koefisien regresi non-linear:

$$\boxed{\boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}}}$$

Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{11}^2 & \dots & x_{11}^n \\ 1 & x_{21} & x_{21}^2 & \dots & x_{21}^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m1}^2 & \dots & x_{m1}^n \end{bmatrix}; \overline{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

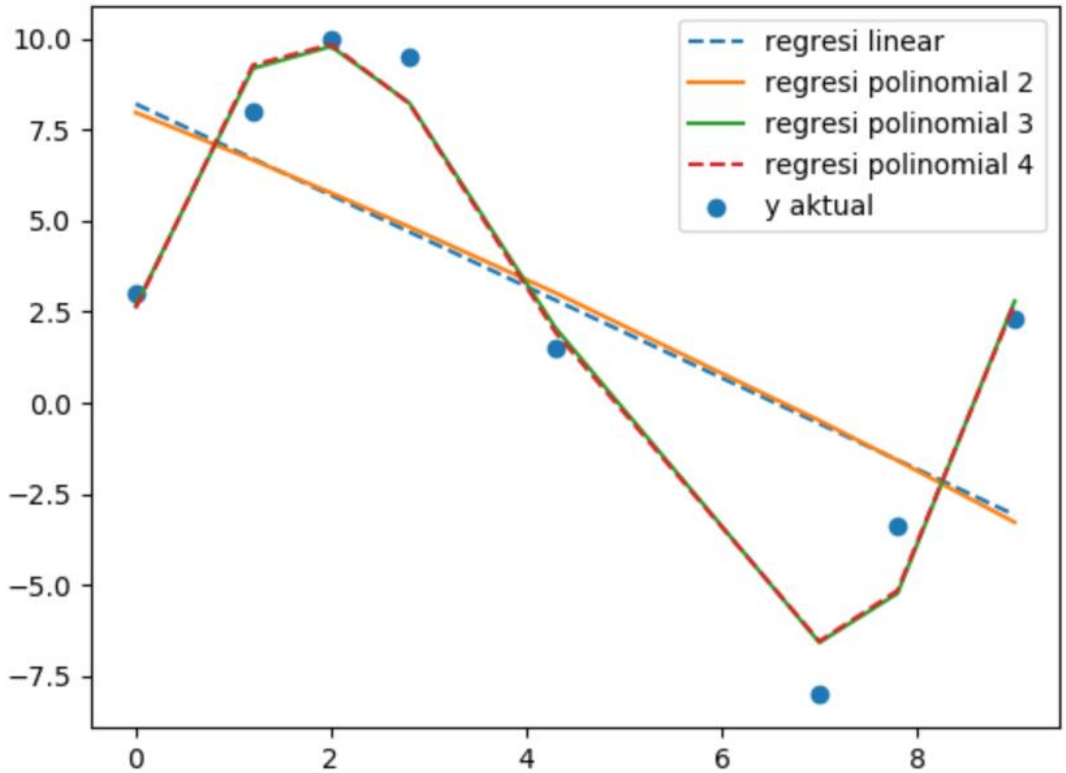
- Apa bedanya?

Latihan

Diberikan data berikut, temukan: (i) model regresi linear dan (ii) regresi polinomial orde 2!

Input data x	Output data Y
0	3
1,2	8
2	10
2,9	9,5
4,3	1,5
7	-8
7,8	-3,4
9,2	2,3

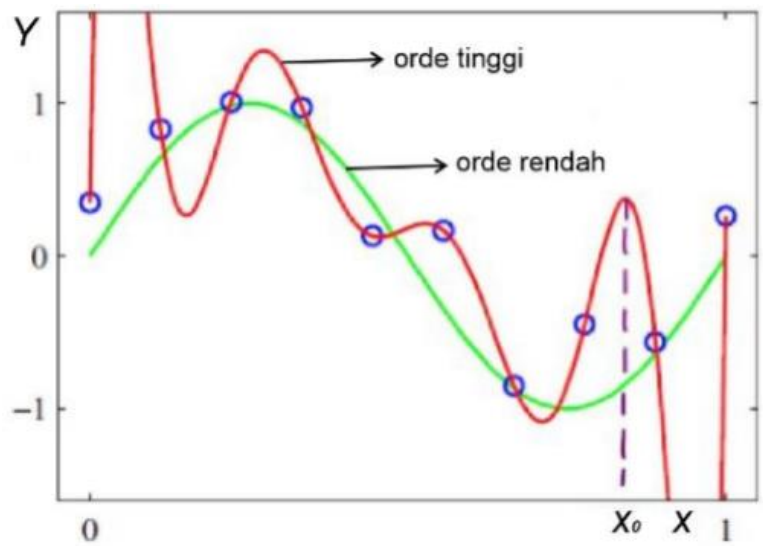
Plot model regresi



Regresi dengan Regularisasi

Regresi dengan Regularisasi

- Di regresi polinomial, semakin tinggi ordernya, model semakin dapat "*fit*" terhadap datanya
- Ingat, hal tersebut dapat menyebabkan *overfitting*.



- Untuk menghindari *overfitting* → regularisasi
  - ✓ **Pendekatan**: koefisien regresi yang kecil, lebih aman terhadap *overfitting*.

Regresi dengan Regularisasi

- Untuk menghindari *overfitting* → regularisasi
  - ✓ **Pendekatan**: koefisien regresi yang kecil, lebih aman terhadap *overfitting*.

Loss function → $f(\boldsymbol{a}) = \sum_{i=1}^{m} (y_i - \bar{y}_i)^2 + \boxed{\lambda \sum_{j=1}^{n} (a_j^2)}$

Untuk mendapatkan koefisien regresi yang kecil

  - ✓ Semakin tinggi λ, semakin "teregularisasi" (model semakin general)
- Formula akhir untuk mendapatan koefisien regresi $\boldsymbol{a}$:

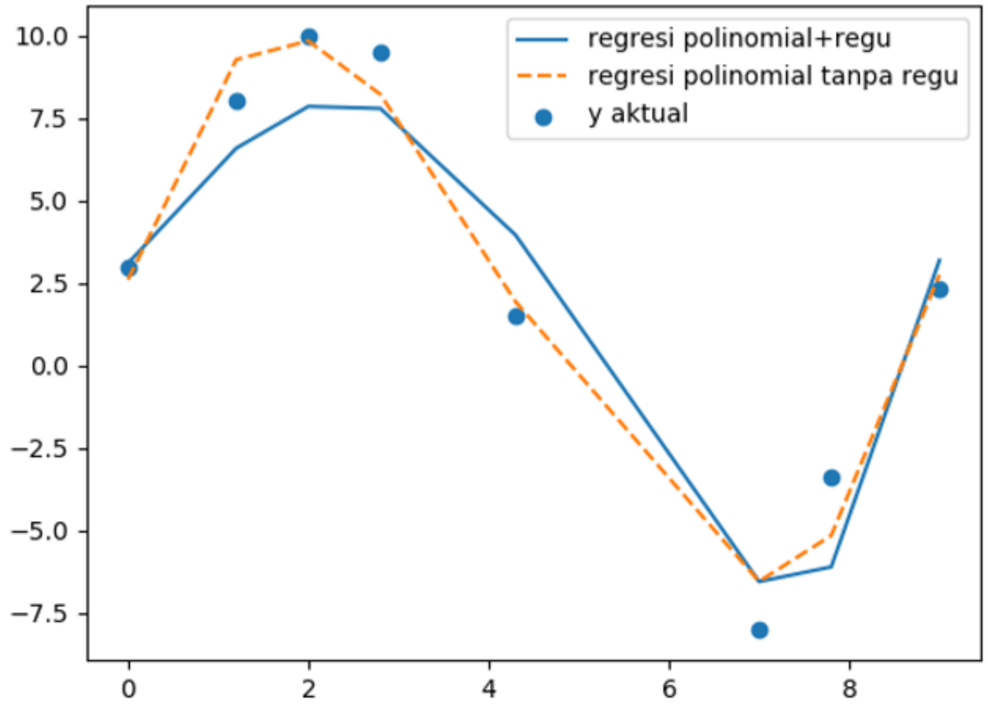
$$\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}}$$

Dengan $\boldsymbol{I}$ adalah matriks identitas berukuran $(n+1, n+1)$
$n =$ orde polinomial

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Regresi dengan Regularisasi

Input data x	Output data Y
0	3
1,2	8
2	10
2,9	9,5
4,3	1,5
7	-8
7,8	-3,4
9,2	2,3



Perbandingan Regresi:
(i) Linear, (ii) Non-linear, (iii) dengan regularisasi

- Formula koefisien regresi linear:

$$\boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}}$$

Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1n} \\ 1 & x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix}; \overline{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

- Formula koefisien regresi non-linear:

$$\boldsymbol{a} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}}$$

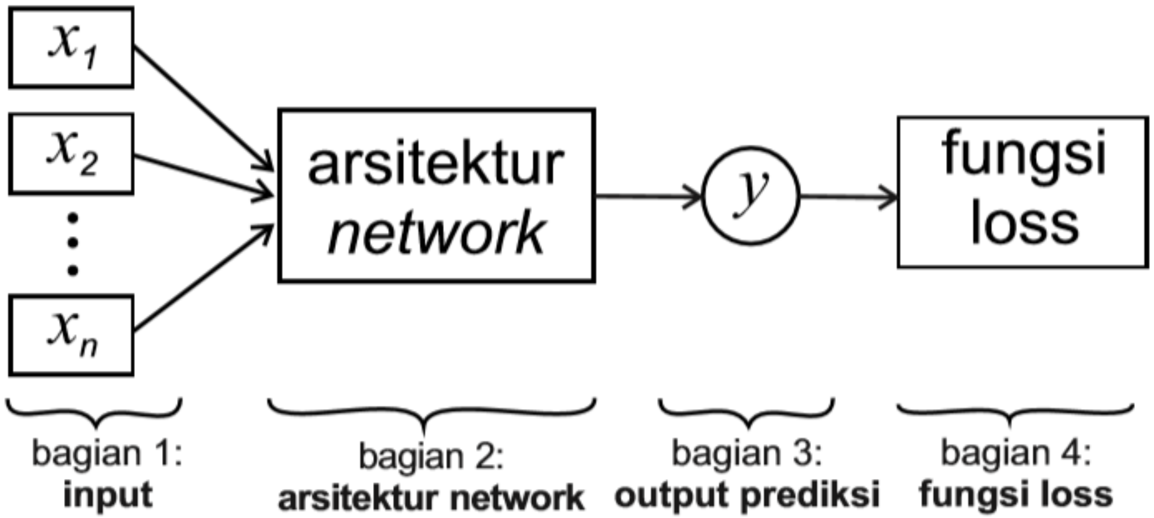
Dengan:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{11}^2 & \dots & x_{11}^n \\ 1 & x_{21} & x_{21}^2 & \dots & x_{21}^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{m1} & x_{m1}^2 & \dots & x_{m1}^n \end{bmatrix}; \overline{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix}$$

- Formula koefisien regresi dengan regularisasi:

$$\boldsymbol{a} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\overline{\boldsymbol{y}}$$

ANN untuk Regresi



Fungsi loss:

$$MSE = \frac{\sum_{i=1}^{m} (\overline{y} - y)^2}{m}$$

di mana :

$\overline{y}$ = label *ground truth*

y = output prediksi

m = banyaknya nilai regresi yang diprediksi

Meminimalkan MSE

Teknik-teknik Regresi Lain

- SVR (*Support Vector Regression*)
- Gaussian processes
- Bayesian regression
- Decision tree regressor

End..