

# Regresi Linear

---

Ali Akbar Septiandri

November 14, 2018

Universitas Al Azhar Indonesia

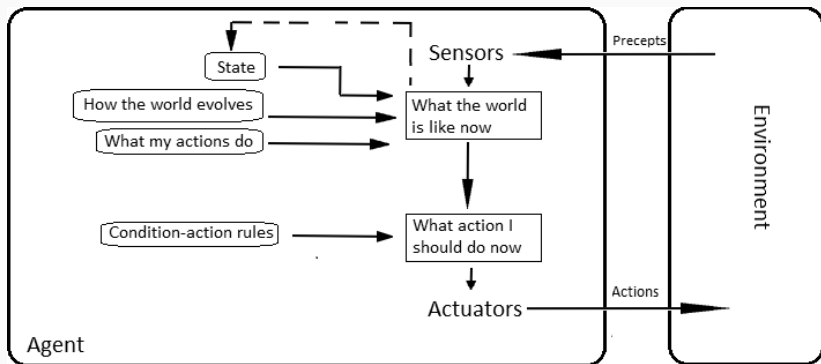
1. Ordinary Least Squares
2. Error Minimisation

**"The only stupid question is the one you were  
afraid to ask but never did."**

**- Richard Sutton**

Office hours dimulai pekan ini.  
Setiap hari Rabu, pukul 08.00-09.00.

# Model-based reflex agents



**Figure 1:** Model yang mengandalkan riwayat persepsi dan dampaknya

# Ordinary Least Squares

---

## Contoh Kasus 1

Anda punya 3 data transaksi taksi yang terdiri dari jarak (km) dan tarif (Rp) sebagai berikut:

jarak (km)	tarif (Rp)
2	12,000
4	18,000
6	24,000

Jika diketahui bahwa tarif dasarnya adalah Rp 6,000, berapa tarif taksi tersebut per km?

## Contoh Kasus 2

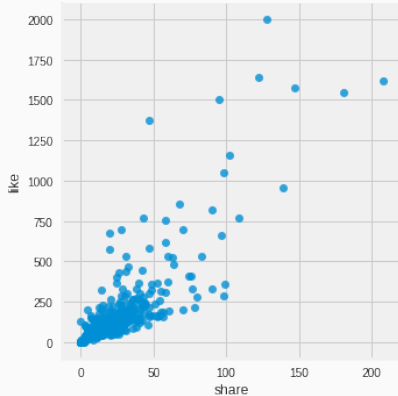
Teman Anda punya 3 data transaksi taksi yang terdiri dari jarak (km) dan tarif (Rp) sebagai berikut:

jarak (km)	tarif (Rp)
2	12,300
4	17,700
6	23,550

Jika diketahui bahwa tarif dasarnya adalah Rp 6,000, berapa tarif taksi tersebut per km?

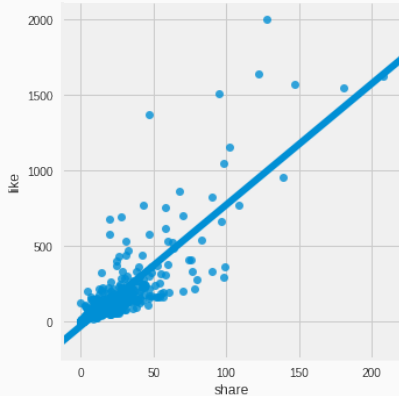


## Prediksi hubungan antara dua variabel



**Figure 2:** Data hubungan antara 'share' dengan 'like' pada Facebook

## Prediksi hubungan antara dua variabel



**Figure 2:** Data hubungan antara 'share' dengan 'like' pada Facebook

# Simple linear regression

## Fungsi linear

Kasus paling sederhana adalah mencocokkan garis lurus ke sekumpulan data

$$y = ax + b$$

dengan  $a$  adalah *slope*, gradien, atau kemiringan; sedangkan  $b$  dikenal dengan nama *intercept* atau bias.

## Notasi lain

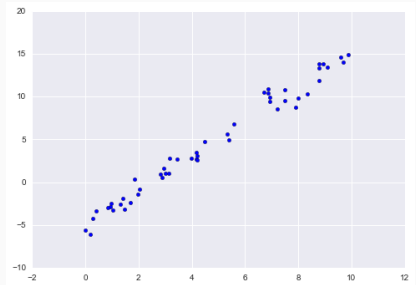
$$y = w_0 + w_1x_1$$

dengan  $w$  adalah bobot atau koefisien.

# Linear regresi dari fungsi yang diketahui

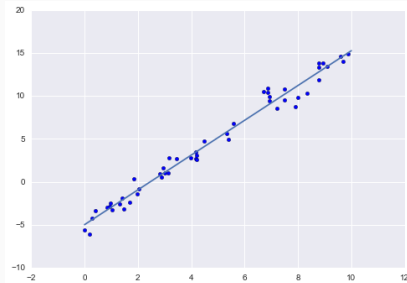
## Example

```
rng = np.random.RandomState(1)
x = 10 * rng.rand(50)
y = 2 * x - 5 + rng.randn(50)
plt.scatter(x, y);
```



**Figure 3:** Data yang dimunculkan secara acak [VanderPlas, 2016]

# Ordinary least squares (OLS) regression



**Figure 4:** Hasil pencocokan garis [VanderPlas, 2016]

Model slope: 2.02720881036

Model intercept: -4.99857708555

Bagaimana kalau ada lebih dari dua variabel  
yang ingin kita lihat hubungannya?

# Multidimensional linear regression

## Model

$$y = w_0x_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D = \sum_{j=0}^D w_jx_j$$

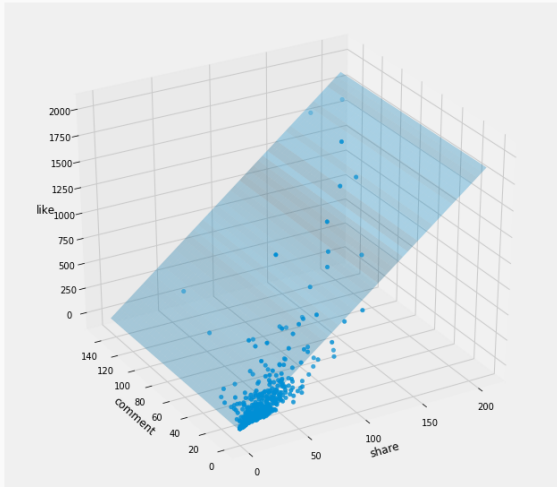
dengan  $x_0 = 1$

## Notasi matriks-vektor

$$y = \mathbf{w}^T \mathbf{x}$$

dengan  $\mathbf{x}$  adalah vektor fitur (*feature vector*)

# Regresi linear untuk dua variabel



**Figure 5:** Hubungan antara 'share', 'comment', dan 'like' pada foto di Facebook



## Prediktor linear (contoh)

Vektor bobot  $\mathbf{w} \in \mathbb{R}^D$

bias: -20.24

share: 6.65

comment: 3.53

Vektor fitur  $\mathbf{x} \in \mathbb{R}^D$

bias: 1

share: 147

comment: 58

$$\hat{y} = \mathbf{w}^T \mathbf{x}$$

$$= \sum_{j=0}^D w_j x_j$$

$$= -20.24(1) + 6.65(147) + 3.53(58) = 1162.05$$

Jadi, *diprediksi* bahwa untuk foto dengan *share* = 147 dan *comment* = 58, foto tersebut akan mendapatkan  $\approx 1162.05$  *likes*.

**Bagaimana cara mendapatkan nilai  $w$ ?**

# Error Minimisation

---

# (Supervised) Learning

- Kita ingin mencari  $f : \mathcal{X} \rightarrow \mathcal{Y}$

$\mathcal{X}$ : data masukan

$\mathcal{Y}$ : data keluaran

dari data latih yang i.i.d.<sup>1</sup>

$$\mathcal{D} = (x_1, y_1), \dots, (y_N, y_N)$$

- Objektif: Meminimalkan *generalisation error* dengan menggunakan *loss function*  $\ell$ , contohnya:

$$\ell(y, f(x)) = (y - f(x))^2$$

yang juga dikenal dengan nama *squared loss*

---

<sup>1</sup>independent and identically distributed

# Meminimalkan error pada data latih

- Untuk meminimalkan *generalisation error*

$$w^* = \arg \min_w \mathbb{E}_{X,Y}[(Y - Xw)^2]$$

# Meminimalkan error pada data latih

- Untuk meminimalkan *generalisation error*

$$w^* = \arg \min_w \mathbb{E}_{X,Y}[(Y - Xw)^2]$$

- Kita tidak punya data untuk seluruh kemungkinan pasangan nilai  $X$  dan  $Y$ !

# Meminimalkan error pada data latih

- Untuk meminimalkan *generalisation error*

$$w^* = \arg \min_w \mathbb{E}_{X,Y}[(Y - Xw)^2]$$

- Kita tidak punya data untuk seluruh kemungkinan pasangan nilai  $X$  dan  $Y$ !
- Ide: Minimalkan error pada data latih

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} \frac{1}{|\mathcal{D}_{train}|} \sum_{(x,y) \in \mathcal{D}_{train}} \ell(y, x, \mathbf{w})$$

- Didefinisikan fungsi error

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

atau

$$E(\mathbf{w}) = \frac{1}{N} (\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w})$$



# Optimasi analitis

- Didefinisikan fungsi error

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

atau

$$E(\mathbf{w}) = \frac{1}{N} (\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w})$$

- Nilainya dapat dioptimasi dengan mencari turunan pertama, lalu atur nilainya menjadi 0, i.e.  $\frac{\partial E}{\partial \mathbf{w}} = 0$  atau  $\nabla_{\mathbf{w}} E(\mathbf{w}) = 0$

# Optimasi analitis

- Didefinisikan fungsi error

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

atau

$$E(\mathbf{w}) = \frac{1}{N} (\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w})$$

- Nilainya dapat dioptimasi dengan mencari turunan pertama, lalu atur nilainya menjadi 0, i.e.  $\frac{\partial E}{\partial \mathbf{w}} = 0$  atau  $\nabla_{\mathbf{w}} E(\mathbf{w}) = 0$
- Solusi tertutupnya:

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$$

- Didefinisikan fungsi error

$$E(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^N (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

atau

$$E(\mathbf{w}) = \frac{1}{N} (\mathbf{y} - X\mathbf{w})^T (\mathbf{y} - X\mathbf{w})$$

- Nilainya dapat dioptimasi dengan mencari turunan pertama, lalu atur nilainya menjadi 0, i.e.  $\frac{\partial E}{\partial \mathbf{w}} = 0$  atau  $\nabla_{\mathbf{w}} E(\mathbf{w}) = 0$
- Solusi tertutupnya:

$$\hat{\mathbf{w}} = (X^T X)^{-1} X^T \mathbf{y}$$

- Bagian  $(X^T X)^{-1} X^T$  dikenal sebagai *pseudo-inverse*

**Bagaimana kalau atributnya bersifat kategori?**

Buatlah model untuk memprediksi *jumlah likes* yang akan didapatkan sebuah foto jika diketahui *usia* pembuat pos, *jenis kelaminnya*, dan *kategori gambarnya* (pemandangan, orang, benda).

- usia =  $\{x_1 \in \mathbb{N}\}$

- $\text{usia} = \{x_1 \in \mathbb{N}\}$
- $\text{jenis kelamin} = \{\text{laki-laki, perempuan}\} = \{0, 1\}$

- usia =  $\{x_1 \in \mathbb{N}\}$
- jenis kelamin =  $\{\text{laki-laki, perempuan}\} = \{0, 1\}$
- kategori =  $\{\text{pemandangan, orang, benda}\} = \{0, 1, 2\}$ ?



- $\text{usia} = \{x_1 \in \mathbb{N}\}$
- $\text{jenis kelamin} = \{\text{laki-laki, perempuan}\} = \{0, 1\}$
- $\text{kategori} = \{\text{pemandangan, orang, benda}\} = \{0, 1, 2\}$ ?
- Apakah  $\text{benda} > \text{pemandangan}$ ?

- $\text{usia} = \{x_1 \in \mathbb{N}\}$
- $\text{jenis kelamin} = \{\text{laki-laki, perempuan}\} = \{0, 1\}$
- $\text{kategori} = \{\text{pemandangan, orang, benda}\} = \{0, 1, 2\}$ ?
- Apakah  $\text{benda} > \text{pemandangan}$ ?
- Apakah  $\text{perempuan} > \text{laki-laki}$ ?

# One-of-k encoding

- Dikenal juga dengan nama “one-hot encoding”

# One-of-k encoding

- Dikenal juga dengan nama “one-hot encoding”
- Mengubah masing-masing nilai dari suatu atribut menjadi atribut tersendiri

# One-of-k encoding

- Dikenal juga dengan nama “one-hot encoding”
- Mengubah masing-masing nilai dari suatu atribut menjadi atribut tersendiri
- e.g. kategori = {pemandangan, orang, benda} menjadi
  - kategori\_pemandangan = {0, 1}
  - kategori\_orang = {0, 1}
  - kategori\_benda = {0, 1}
- sehingga...

$$y = w_0x_0 + w_1x_1 + \dots + w_6x_6 = \sum_{j=0}^6 w_jx_j$$

dengan  $x_0 = 1$ ,  $x_1 = \text{usia}$ ,  $x_2 = \text{jk\_laki}$ ,  $x_3 = \text{jk\_perempuan}$ ,  
 $x_4 = \text{kategori\_pemandangan}$ ,  $x_5 = \text{kategori\_orang}$ , dan  
 $x_6 = \text{kategori\_benda}$ .

- Regresi linear dapat digunakan untuk **memprediksi nilai riil**
- Regresi linear mempunyai **solusi tertutup** untuk mencari nilai bobot
- Gunakan **one-hot encoder** untuk mengubah atribut bertipe kategori

## Pertemuan berikutnya

- Klasifikasi: regresi logistik
- Optimasi numerik





Jake VanderPlas (2016)

## **In Depth: Linear Regression**

*Python Data Science Handbook*

Terima kasih