



IPB University
— Bogor Indonesia —

Department of Statistics
Faculty of Mathematics and Natural Sciences



STA1381 Pengantar Sains Data

Pengenalan Staistical Machine Learning
Supervised Learning

Anang Kurnia
Departemen Statistika, FMIPA - IPB
[anangk\[at\]apps.ipb.c.id](mailto:anangk[at]apps.ipb.c.id)

Rencana Perkuliahan Sesi UAS:

Minggu	Materi
8	Nonlinear Regression (1)
9	Nonlinear Regression (2)
10	Pengenalan Stat. Machine Learning: Supervised Learning
11	Pengenalan Stat. Machine Learning: Seleksi Peubah
12	Pengenalan Stat. Machine Learning: Unsupervised Learning
13	Aplikasi
14	Presentasi kelompok
UAS	

What is Statistical Learning?

- Suppose we observe Y_i and $X_i = (X_{i1}, X_{i2}, \dots, X_{ip})$ for $i = 1, 2, \dots, n$.
- We believe that there is a relationship between Y and at least one of the X 's.
- We can model the relationship as

$$Y_i = f(\mathbf{X}_i) + \varepsilon_i$$

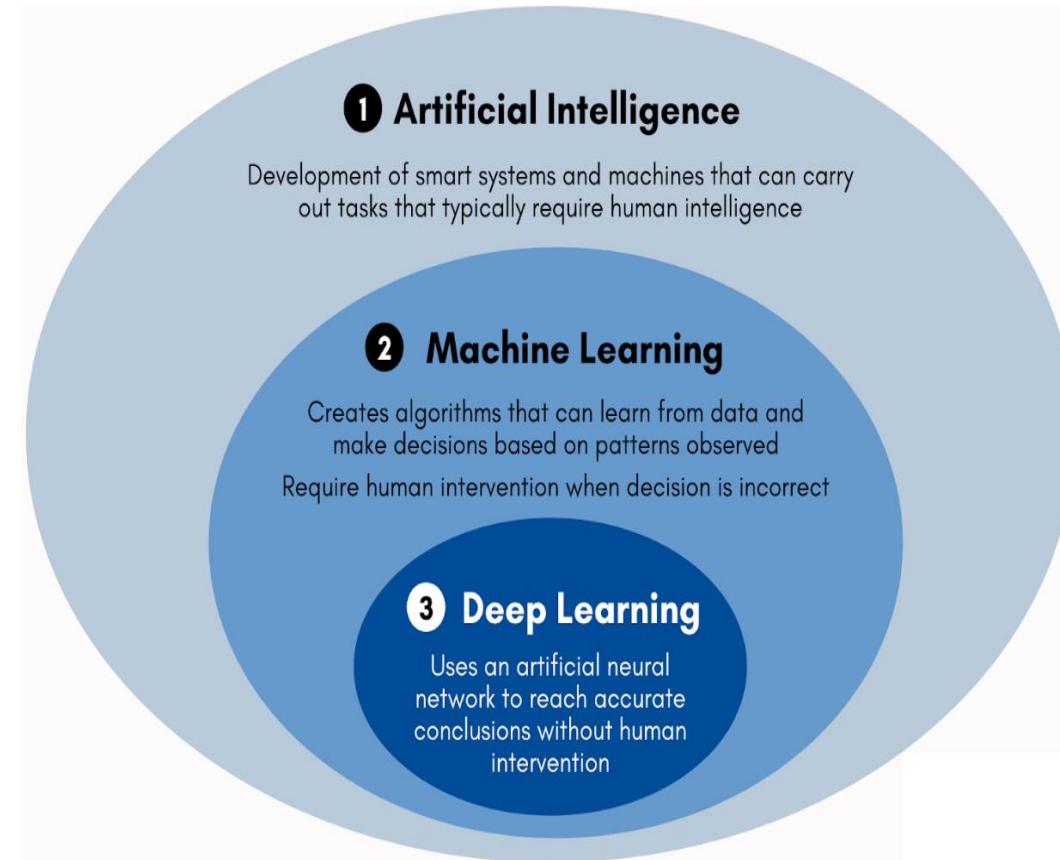
where f is an unknown function and ε is a random error with mean zero.

Outline

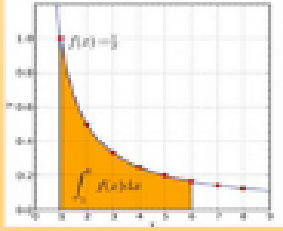
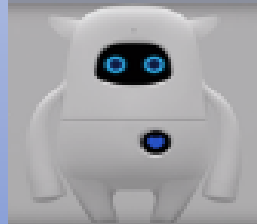
- Pengantar
- Statistical Learning
 - Supervised vs unsupervised
- Beberapa metode supervised
 - Ridge Regression
 - Lasso Regression
 - Model Averaging

Pengantar

- Tujuan machine learning secara umum adalah untuk memahami struktur data dan memasukkan data tersebut ke dalam model yang dapat dipahami dan dimanfaatkan.
- Machine learning \leftarrow membangun model dari data contoh untuk “mengotomatisasi” proses pengambilan keputusan berdasarkan input data.
- Machine learning \leftarrow dapat menggunakan analisis statistik untuk menghasilkan nilai sesuai tujuan analisisnya.

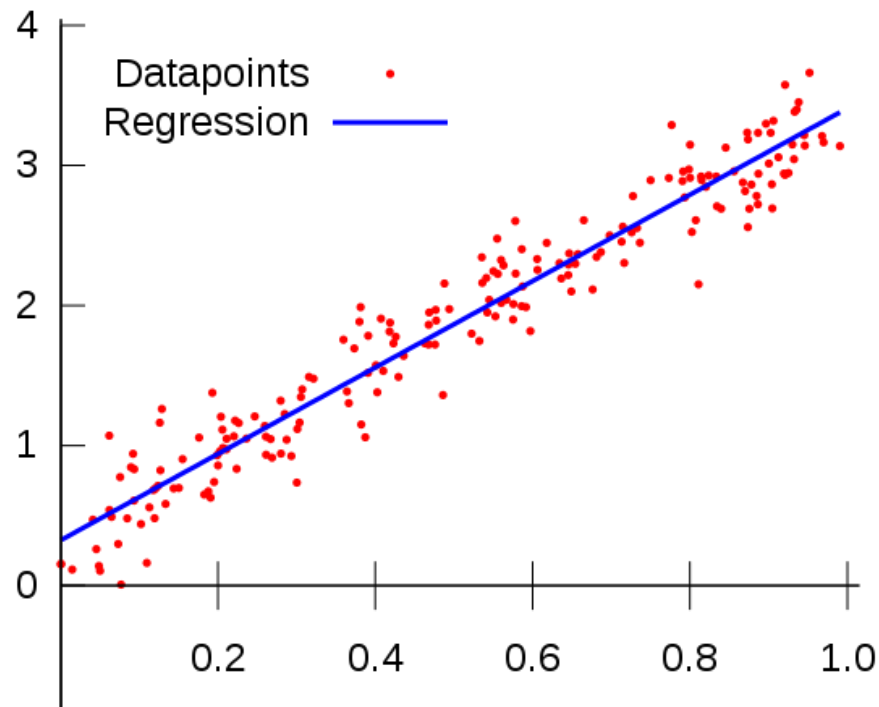


Machine learning vs statistika (statistics)?

	 Statistics	 Machine Learning
Subfield of...	Mathematics	Computer Science (AI)
Focus on...	Building models with explicitly programmed instructions	Creating systems that learn from data
Purpose	Inferences; Relationships between variables	Optimization; Prediction accuracy
Prior assumptions about data	Some knowledge about population usually required	None
Dimensionality of data	Usually applied to low-dimensional data	Usually applied to high dimensional data; ML learns from data
Knowledge overlap	No ML knowledge required	Some stats knowledge usually needed; stats is basis for algorithms

Ilustrasi 1

Statistical Models vs Machine learning — Linear Regression Example



Machine Learning:

- Tujuan machine learning, dalam hal ini, adalah untuk mendapatkan performa terbaik pada set pengujian.
- Sehingga prakteknya dibentuk **data training untuk pembentukan model**, dan **data testing untuk evaluasi model** tersebut

Statistical Models:

- Menemukan garis linear yang meminimalkan JKG, dengan asumsi Gaussian, dan **tidak ada data training dan testing yang diperlukan**.
- Tujuannya lebih untuk identifikasi peubah prediktor yang mempengaruhi peubah respon, meskipun juga dapat digunakan untuk prediksi

Ilustrasi 2

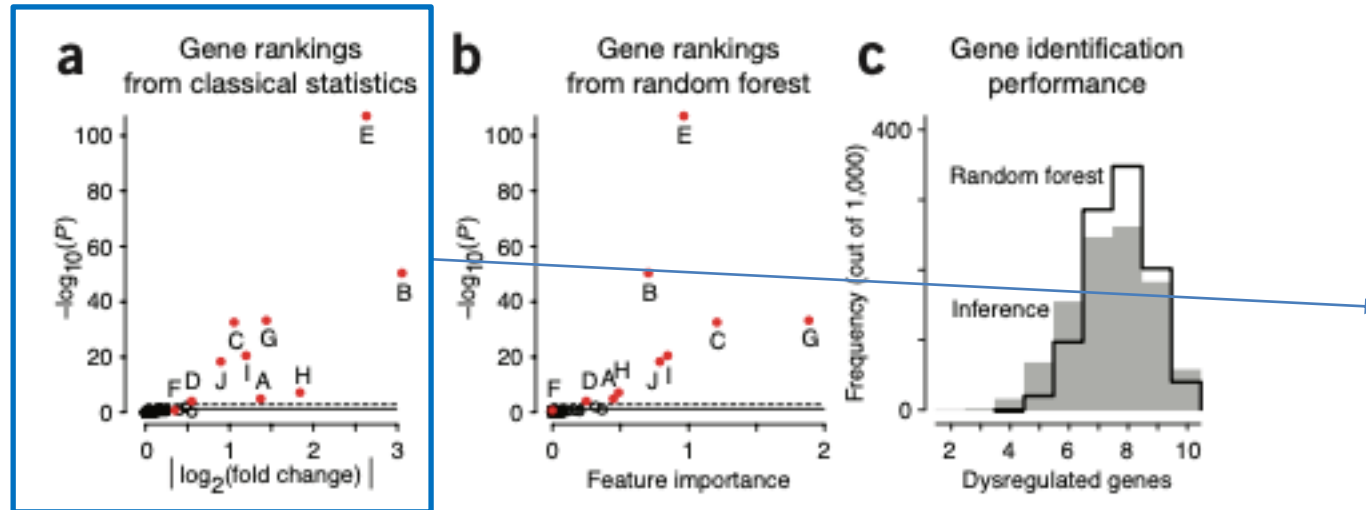
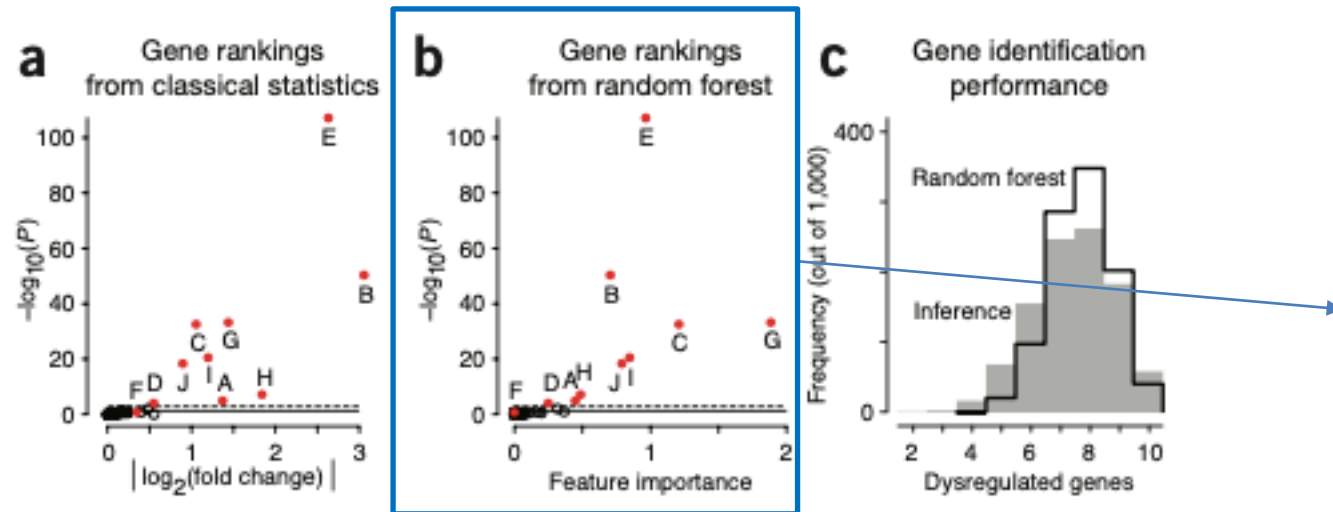


Figure 2 | Analysis of gene ranking by classical inference and ML. (a) Unadjusted log-scaled P values from statistical differential expression analysis as a function of effect size, measured by fold change in expression. (b) Log-scaled P values from a as a function of gene importance from random forest classification. In a and b, red circles identify the ten differentially expressed genes from **Figure 1**; the remaining genes are indicated by open circles. (c) Distribution of the number of dysregulated genes correctly identified in 1,000 simulations by inference (gray fill) and random forest (black line).

menunjukkan p-value dari uji antar fenotipe sebagai fungsi dari perubahan log fold dalam ekspresi gen. Sepuluh gen disregulasi ditampilkan dengan titik berwarna merah; Hasilnya: diperoleh sembilan dari sepuluh (kecuali F, dengan perubahan log fold terkecil) sebagai gen yang signifikan dengan p-value $<0,05$.

Ilustrasi 2



hasil klasifikasi RF dengan 100 pohon, dimana p-value dari inferensi klasik (hasil a) diplot sebagai fungsi feature importance (gen). Skor ini mengkuantifikasi kontribusi gen tertentu terhadap peningkatan klasifikasi rata-rata dalam sebuah partisi ketika pohon dipisah memilih gen itu.

Figure 2 | Analysis of gene ranking by classical inference and ML. (a) Unadjusted log-scaled P values from statistical differential expression analysis as a function of effect size, measured by fold change in expression. (b) Log-scaled P values from a as a function of gene importance from random forest classification. In a and b, red circles identify the ten differentially expressed genes from Figure 1; the remaining genes are indicated by open circles. (c) Distribution of the number of dysregulated genes correctly identified in 1,000 simulations by inference (gray fill) and random forest (black line).

Ilustrasi 2

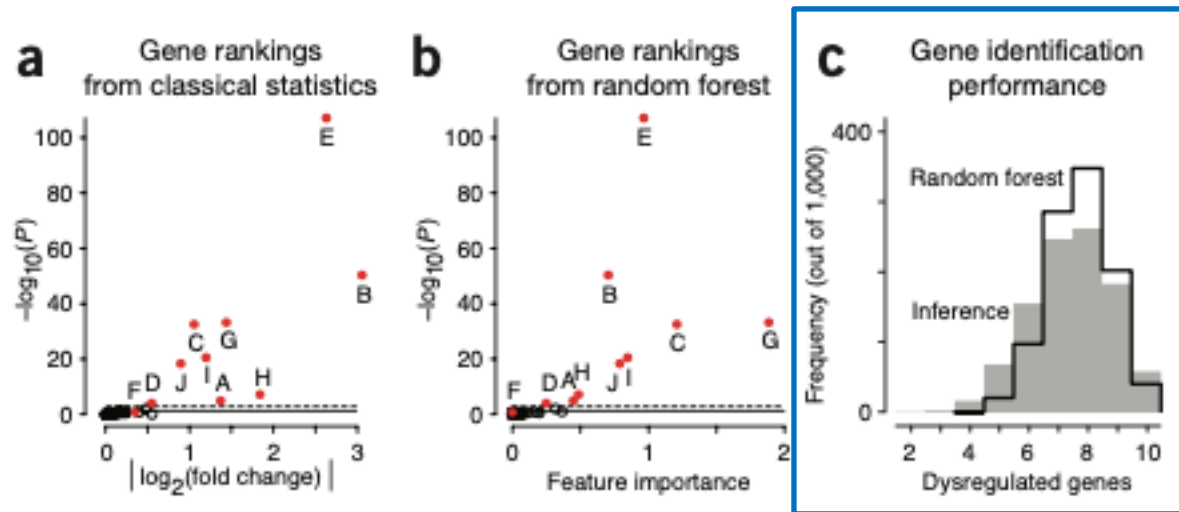


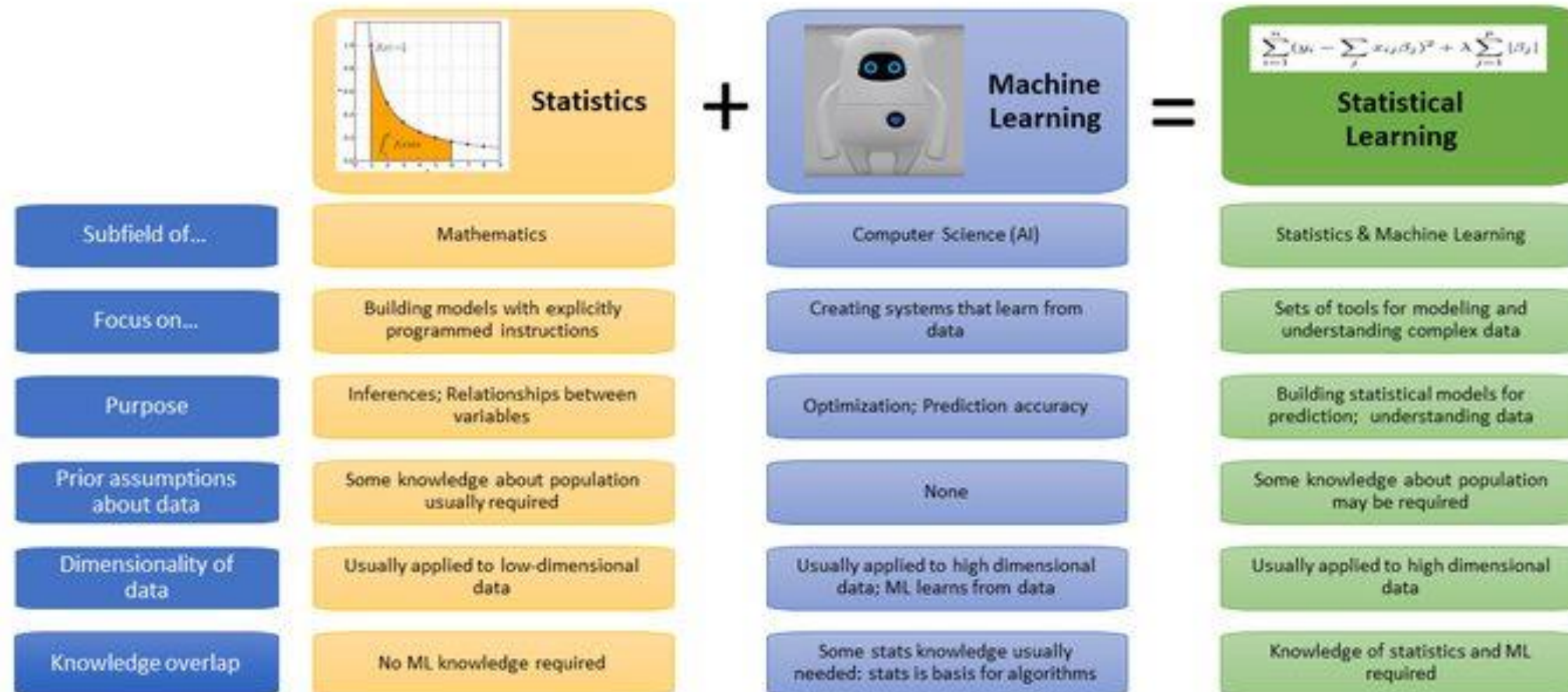
Figure 2 | Analysis of gene ranking by classical inference and ML. (a) Unadjusted log-scaled P values from statistical differential expression analysis as a function of effect size, measured by fold change in expression. (b) Log-scaled P values from a as a function of gene importance from random forest classification. In a and b, red circles identify the ten differentially expressed genes from Figure 1; the remaining genes are indicated by open circles. (c) Distribution of the number of dysregulated genes correctly identified in 1,000 simulations by inference (gray fill) and random forest (black line).

jika kita melakukan simulasi 1.000 kali dan menghitung jumlah gen disregulasi yang diidentifikasi dengan benar oleh kedua pendekatan (a dan b)—berdasarkan uji hipotesis klasik atau generalisasi pola prediktif dengan RF dan sepuluh peringkat feature importance— maka diperoleh bahwa kedua metode menghasilkan hasil yang serupa.

Jumlah rata-rata gen disregulasi yang diidentifikasi adalah 7,4/10 untuk inferensi klasik dan 7,7/10 untuk RF

- Machine learning adalah semua tentang hasil/prediksi/klasifikasi, sedangkan pemodelan statistik lebih tentang menemukan hubungan antara peubah dan signifikansi hubungan tersebut, dan dapat juga untuk menghasilkan prediksi.
- Statistics menarik kesimpulan populasi dari contoh, sedangkan machine learning menemukan pola prediktif yang dapat digeneralisasikan.

Statistics + Machine Learning = Statistical (Machine) Learning



Musio image: Akawikiplc [CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>)]

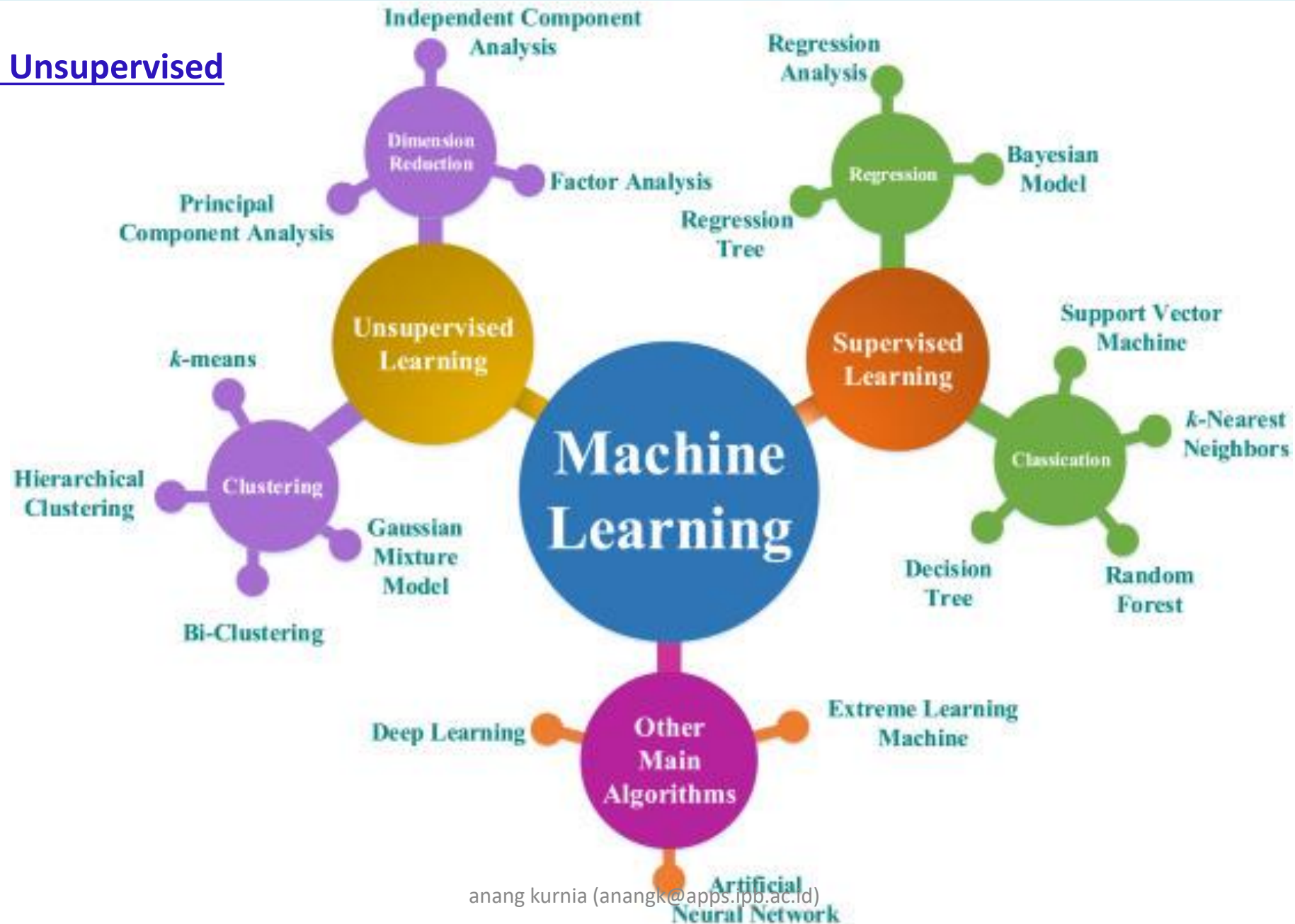
<https://www.datasciencecentral.com/wp-content/uploads/2021/10/3541473617.jpg>

anang kurnia : anangk[at]apps.ipb.ac.id

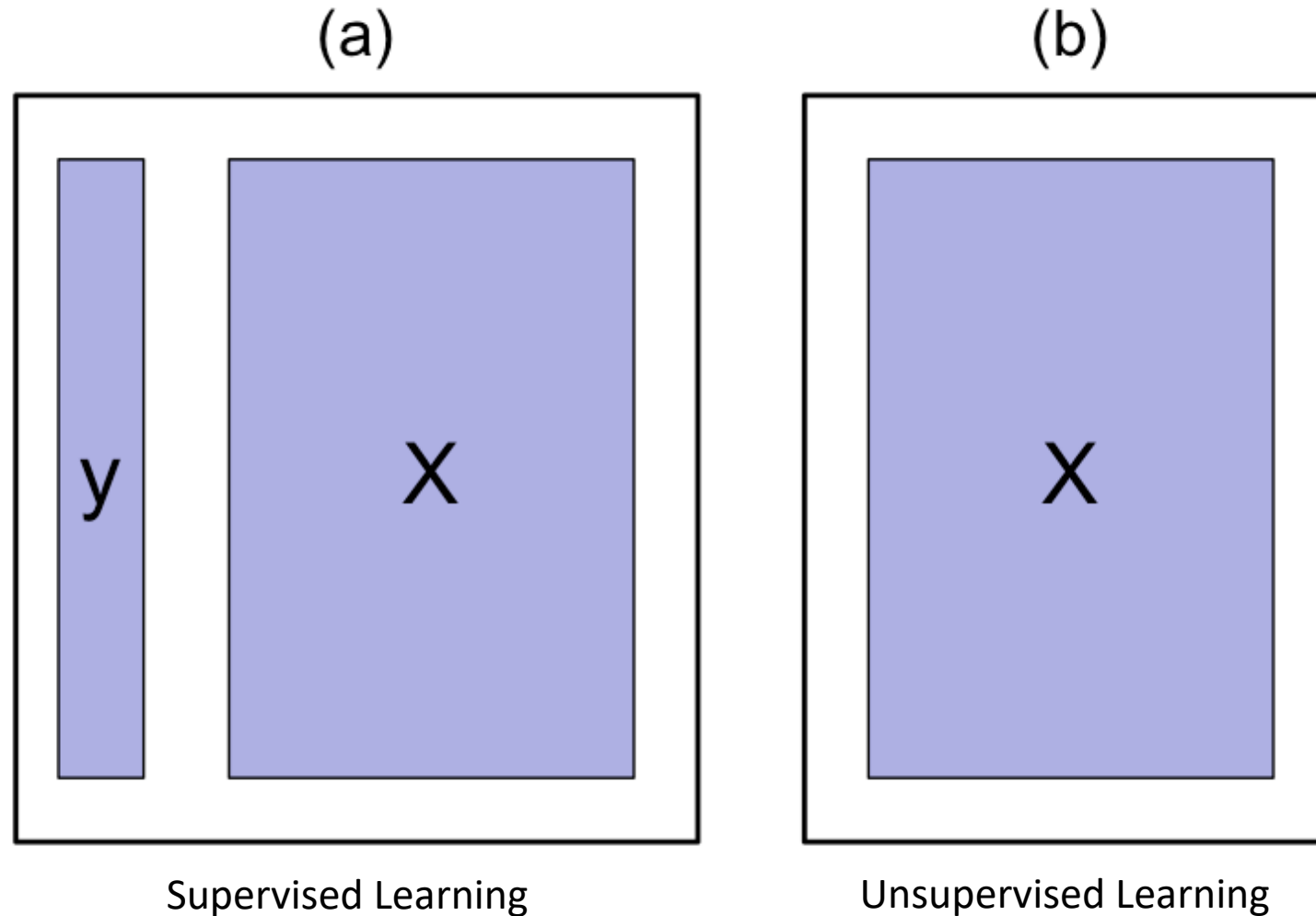
Statistical Learning

- Pembelajaran statistik mengacu pada seperangkat alat yang luas untuk memahami data.
- Alat-alat ini dapat diklasifikasikan sebagai: **supervised learning** dan **unsupervised learning**.
- Secara umum, statistical supervised learning melibatkan pembangunan model statistik untuk memprediksi, atau memperkirakan *outputs* berdasarkan satu atau lebih *inputs*. Masalah dalam ini terjadi di berbagai bidang seperti bisnis, kedokteran, astrofisika, dan kebijakan publik.
- Dengan statistical unsupervised learning, ada *inputs* tetapi tidak ada *supervising outputs*; namun kita dapat mempelajari hubungan dan struktur dari data tersebut.

Supervised vs Unsupervised

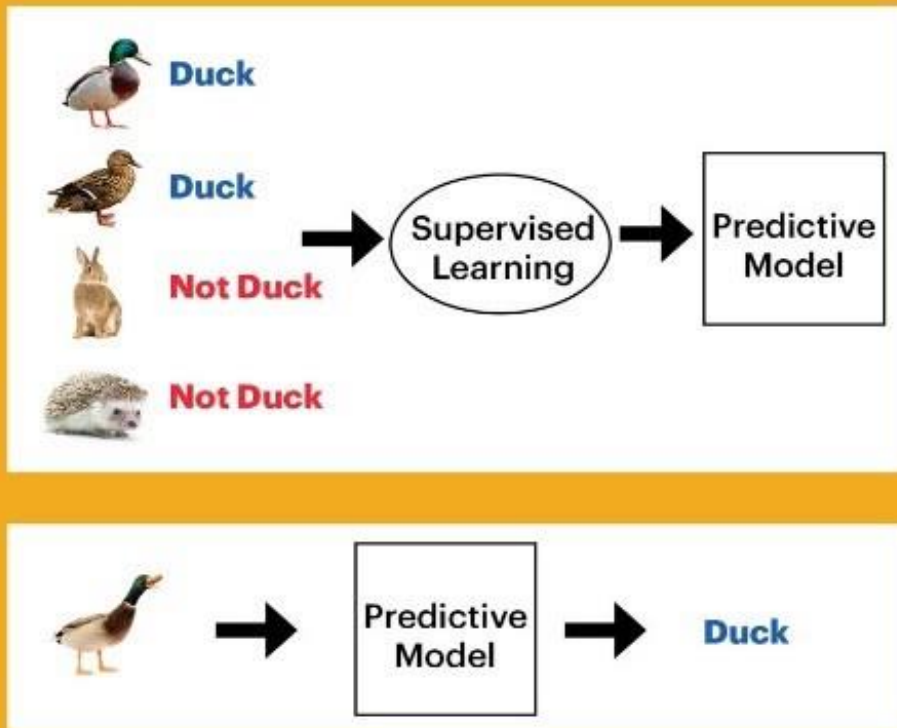


Supervised vs Unsupervised

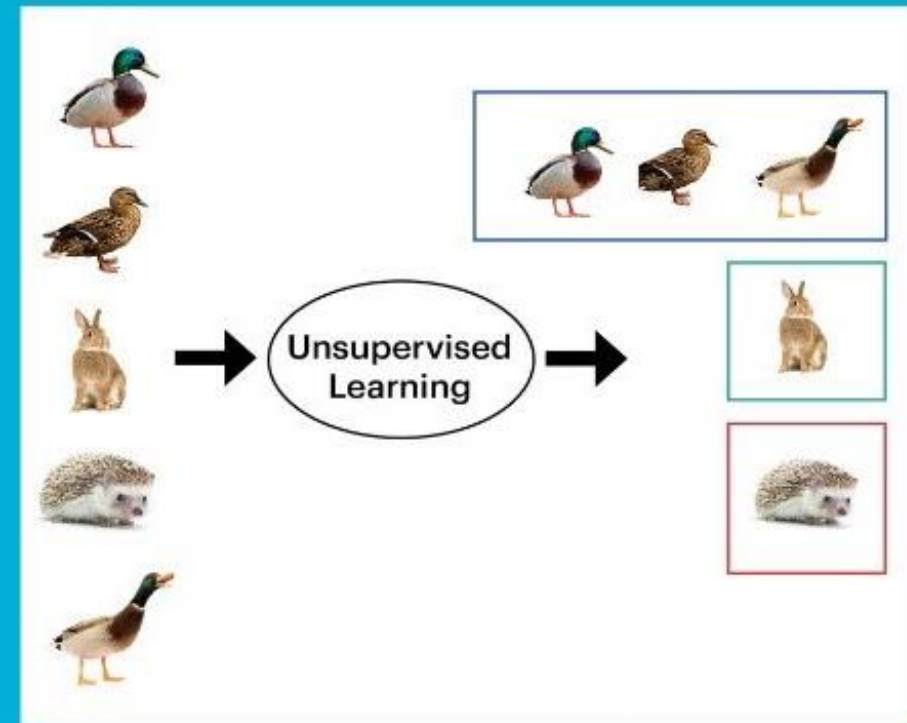


Supervised vs Unsupervised

Supervised Learning (Classification Algorithm)



Unsupervised Learning (Clustering Algorithm)



Beberapa Metode: Supervised Learning

- Supervised Learning:
 - Misalkan diketahui peubah respon Y dengan skala numerik dan beberapa peubah prediktor X_1, X_2, \dots, X_p .
 - Ingin diketahui:
 1. Peubah prediktor mana yang mempengaruhi perubahan rata-rata peubah respon
 2. Prediksi peubah respon
 - Dengan ukuran data yang cukup besar, seringkali metode-metode klasik tidak cocok untuk digunakan.
 - Terdapat metode dalam statistical learning yang dapat digunakan:
 - Ridge regression
 - Lasso Regression
 - Model Averaging

Ridge Regression

Ridge regression (Hoerl & Kennard 1988)

→ meminimalkan jumlah kuadrat galat yang terikat pada regularisasi L_2 dari koefisiennya.

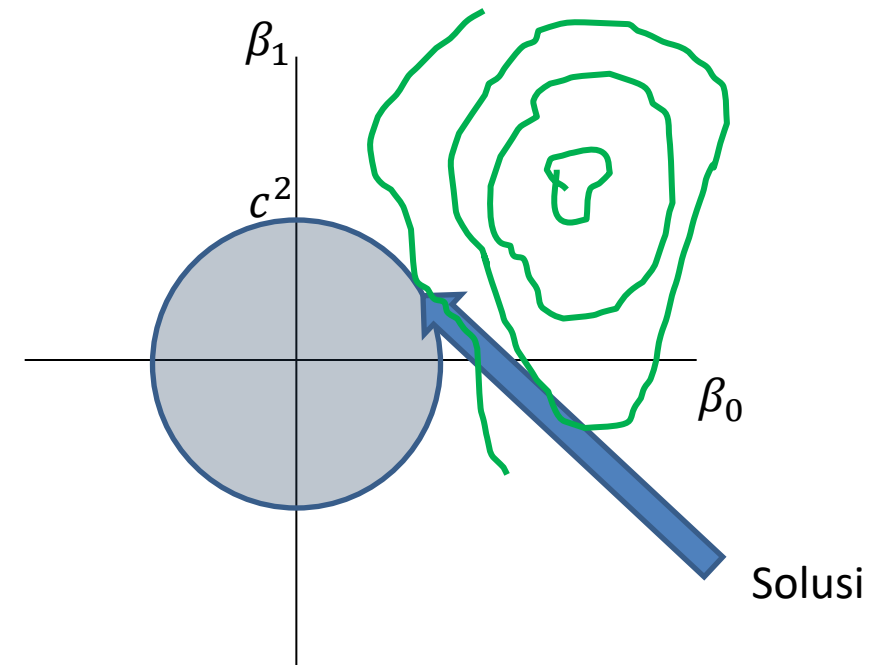
$$\begin{aligned}\hat{\boldsymbol{\beta}}^{Ridge} &= \arg \min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \} \\ &= \arg \min_{\boldsymbol{\beta}} \{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^T \boldsymbol{\beta} \}\end{aligned}$$

$$\Rightarrow \hat{\boldsymbol{\beta}}^{Ridge} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

Pada kasus dengan parameter β_0 and β_1

$$\hat{\boldsymbol{\beta}}^{Ridge} = \arg \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \text{ s.t. } \|\boldsymbol{\beta}\|_2^2 \leq c^2$$

$$\beta_0^2 + \beta_1^2 \leq c^2$$



Kegunaan Ridge Regression

- Menduga koefisien regresi dengan peubah prediktornya saling berkorelasi (multikolinieritas tinggi)
- Digunakan pada pemodelan regresi dengan peubah prediktor sangat banyak (bahkan $p \gg n$)

Sifat dugaan parameteranya:

- Dugaan parameteranya berbias
- Model ridge selalu mempertahankan semua prediktornya

Lasso Regression

(Tibshirani 1996)

→ lasso melakukan penyusutan berkelanjutan (shrinkage) dan pemilihan variabel otomatis secara bersamaan

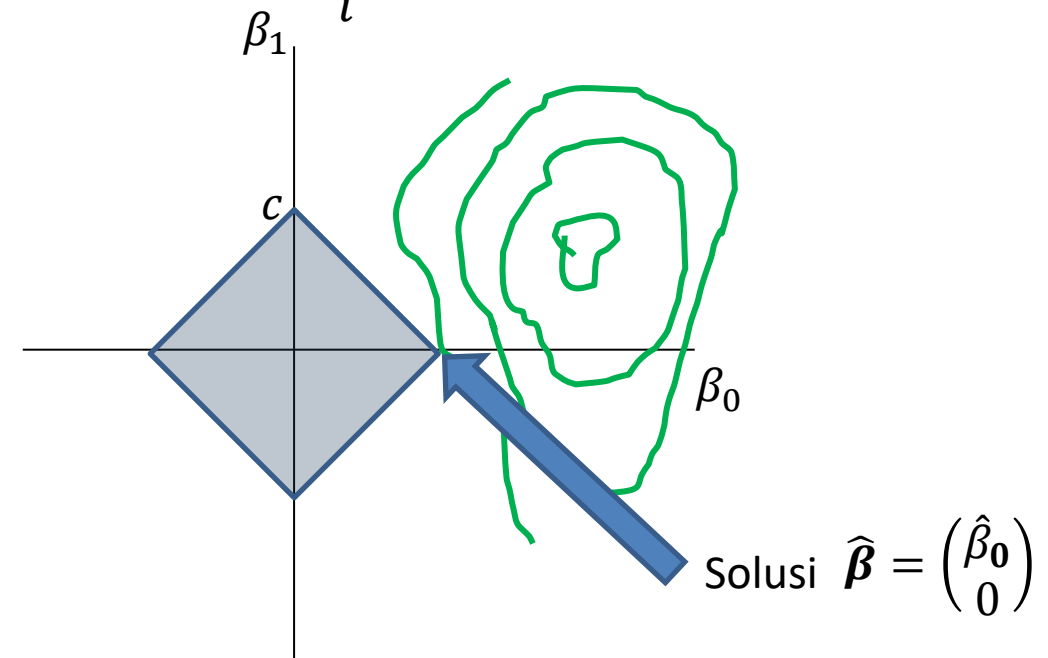
$$\hat{\beta}^{lasso} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_1 \longrightarrow \|\mathbf{a}\|_1 = \sum_i |a_i|$$

➔ *least angle regression* (LARS) algorithm

Pada kasus dengan parameter β_0 and β_1

$$\hat{\beta}^{lasso} = \arg \min_{\beta} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 \text{ s.t. } \|\beta\|_1 \leq c$$

$$|\beta_0| + |\beta_1| \leq c$$



Kegunaan Lasso Regression

- Lasso berguna dalam seleksi peubah prediktor dalam model
- Digunakan pada pemodelan regresi dengan peubah prediktor sangat banyak (bahkan $p \gg n$)

Sifat dugaan parameteranya:

- Dugaan parameteranya berbias
- Pada model lasso memungkinkan tidak semua peubah prediktor dipilih (memiliki koefisien tidak sama dengan 0)

Model Averaging

- Membangun beberapa kandidat model untuk dikombinasikan menjadi sebuah model final
- Umumnya model averaging diterapkan dalam rangka untuk memperoleh nilai prediksi peubah respon

High dimensional regression data ($p \gg n$)

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	...	X_p
y_1	x_{11}	x_{21}	x_{31}	x_{41}	x_{51}	x_{61}	x_{71}	x_{81}		x_{p1}
y_2	x_{12}	x_{22}	x_{32}	x_{42}	x_{52}	x_{62}	x_{72}	x_{82}		x_{p2}
:										
y_n	x_{1n}	x_{2n}	x_{3n}	x_{4n}	x_{5n}	x_{6n}	x_{7n}	x_{8n}		x_{pn}

Model Candidate
Construction

- 1 $Y \sim X_1 + X_3 + X_6 \longrightarrow \hat{Y}_1$
- 2 $Y \sim X_3 + X_4 + X_8 \longrightarrow \hat{Y}_2$
- :
- k $Y \sim X_2 + X_6 + X_p \longrightarrow \hat{Y}_k$

Model Averaging $\hat{Y} = \frac{\sum_{i=1}^k w_i \hat{Y}_i}{\sum_{i=1}^k w_i}$

Ilustrasi di R

```
set.seed(123)
x <- cbind(1,matrix(rnorm(100*20,2,1),100,20))
e <- matrix(rnorm(100),100,1)
b <- c(1,rep(0:4,each=4))
y <- x%*%b+e
```

```
dt.all <- data.frame(y,x[,-1])
str(dt.all)
```

```
#regresi linier
mod1 <- lm(y~.,data=dt.all)
coef(mod1)
```

```
library(glmnet)
#regresi ridge
mod2 <- cv.glmnet(x[,-1],y,alpha=0) #pemilihan lambda dgn cv untuk ridge
mod2
coef(mod2,s="lambda.min")
```

```
#lasso
mod3 <- cv.glmnet(x[,-1],y,alpha=1) #pemilihan lambda dgn cv untuk lambda
mod3
coef(mod3,s="lambda.min")
```

```
#model averaging
library(MuMIn)
mod1 <- lm(y~.,data=dt.all, na.action = na.fail)
mod4 <- dredge(global.model=mod1,m.lim=c(18,20))
mod5 <- model.avg(mod4,delta<4)
mod5
summary(mod5)
```

Bersambung



IPB University

— Bogor Indonesia —

Inspiring Innovation with Integrity
in Agriculture, Ocean and Biosciences for a Sustainable World