Statistika ↔ Machine Learning

Rizal Zaini Ahmad Fathony

Post-doctoral Researcher:
Carnegie Mellon University (CMU) dan *Bosch Center for Artificial Intelligence (BCAI)*

MS & PhD | Computer Science | University of Illinois at Chicago
BS | Komputasi Statistik | Sekolah Tinggi Ilmu Statistik

Statistika vs Machine Learning

Apa itu machine learning? Lagi populer banget sekarang.

Kok mirip dengan Statistika, bedanya apa?

Apa machine learning itu hanya statistik yang di *rebranding* ulang?

Logitic regression itu statistika atau machine learning?

Statistika

Definisi (Wikipedia)

> **Statistics** is the discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

Aplikasi

Ilmu Sosial

Fisika

Bio-informatika

Ekonomi

Demografi

Psikologi

Geologi

Aktuaria

Lingkungan

Bisnis

Biologi

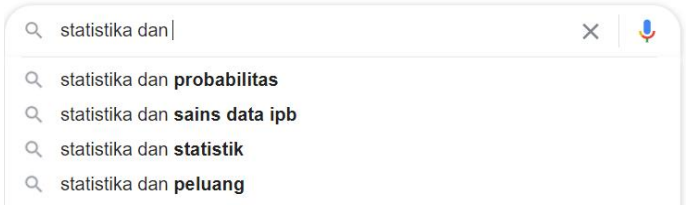
Epidemiologi

Machine Learning

Definisi (Wikipedia)

> **Machine learning** (ML) is the study of computer algorithms that improve automatically through experience. Machine learning algorithms build a model based on sample **data**, in order to make predictions or decisions without being explicitly programmed to do so.

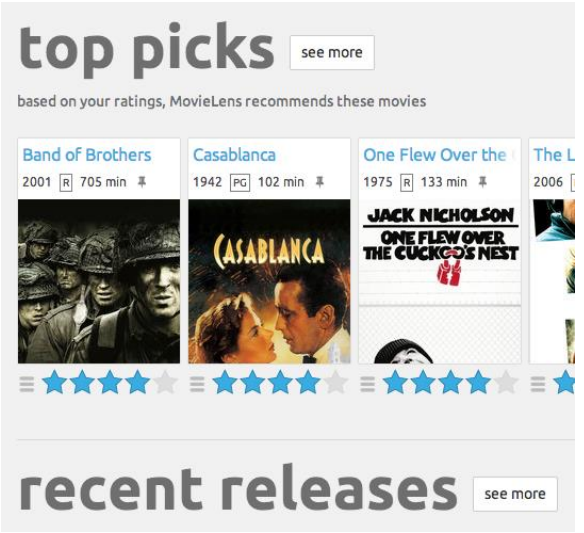
Aplikasi



Web Search

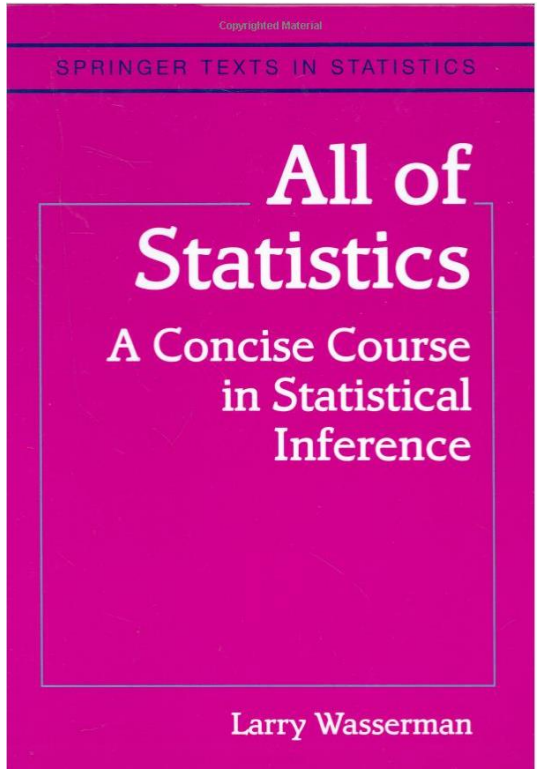
Face Tagging

Spam Filter



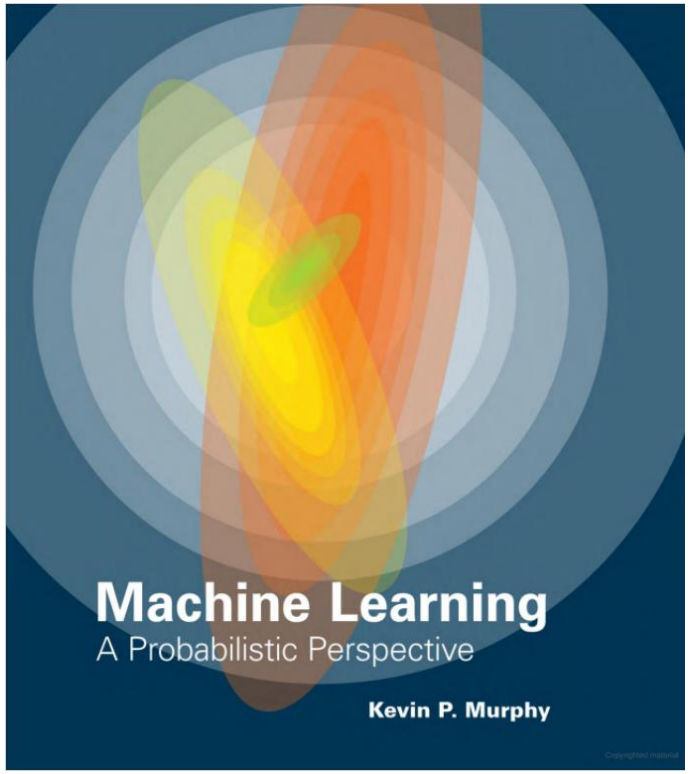
Recommendation

Ilmu Statistika dan Machine Learning



All of Statistics:
A Concise Course in Statistical Inference

Larry Wasserman
Department of Statistics and Data Science
Carnegie Mellon University



Machine Learning
A Probabilistic Perspective

Kevin P. Murphy
Department of Computer Science
University of British Columbia &
Google Research

Topik-topik Statistika dan Machine Learning

All of Statistics: A Concise Course in Statistical Inference

Probability

1. Random Variables
2. Expectation
3. Inequalities
4. Convergence of Random Variables

Statistical Inference

5. Models, Statistical Inference and Learning
6. Estimating the CDF and Statistical Functionals
7. The Bootstrap
8. Parametric Inference
9. Hypothesis Testing and p-values
10. Bayesian Inference
11. Statistical Decision Theory

Statistical Models and Methods

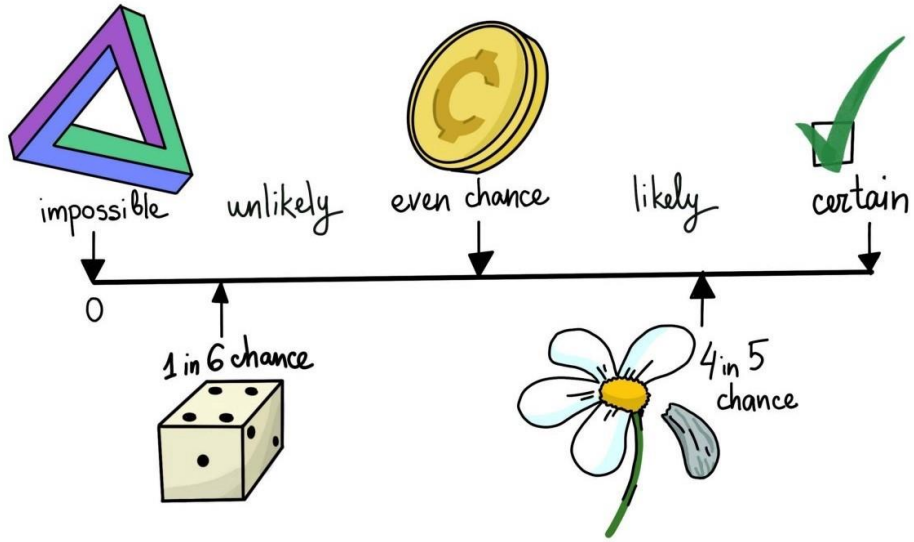
12. Linear and Logistic Regression
13. Multivariate Models
14. Inference About Independence
15. Causal Inference
16. Directed Graphs and Conditional Independence
17. Undirected Graphs
18. Log-Linear Models
19. Nonparametric Curve Estimation
20. Smoothing Using Orthogonal Functions
21. Classification
22. Probability Redux: Stochastic Processes
23. Simulation Methods

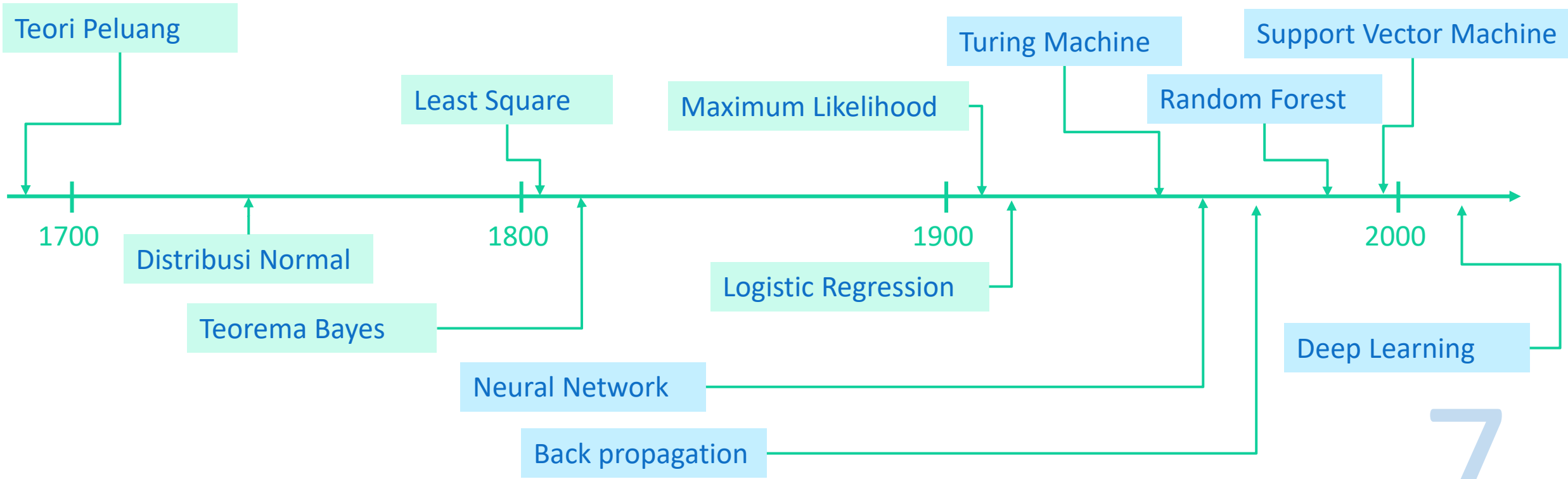
Machine Learning A Probabilistic Perspective

1. Probability
2. Generative models for discrete data
3. Gaussian models
4. Bayesian statistics
5. Frequentist statistics
6. Linear regression
7. Logistic regression
8. Generalized linear models and the exponential family
9. Directed graphical models (Bayes nets)
10. Mixture models and the EM algorithm
11. Latent linear models
12. Sparse linear models
13. Kernels
14. Gaussian processes
15. Adaptive basis function models
16. Markov and hidden Markov models
17. State space models
18. Undirected graphical models (Markov random fields)
19. Exact inference for graphical models
20. Variational inference
21. More variational inference
22. Monte Carlo inference
23. Markov chain Monte Carlo (MCMC) inference
24. Clustering
25. Graphical model structure learning
26. Latent variable models for discrete data
27. Deep learning

Fondasi Utama dan Sejarah

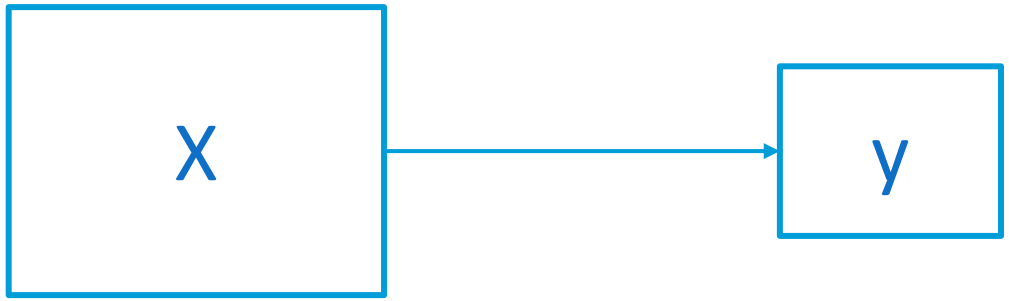
Fondasi Utama: **Teori Peluang**





Statistika vs ML: Contoh Kasus

Analisis Kelulusan Mahasiswa

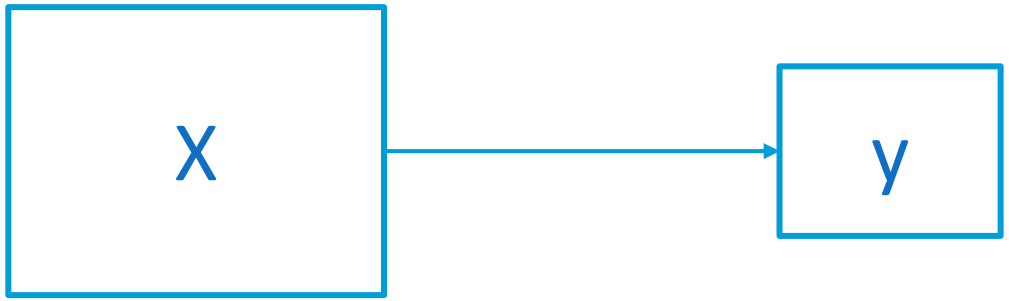


Profil mahasiswa
Nilai mata kuliah
Jam belajar
Aktifitas kampus
Keaktifian di kelas
Jam tidur

Lulus / Tidak Lulus

Logistic Regression

Prediksi Email Spam



Pengirim
Penerima
Alamat email
Server email
Apakah email mengandung attachment
Email mengandung kata ‘lottery’
....

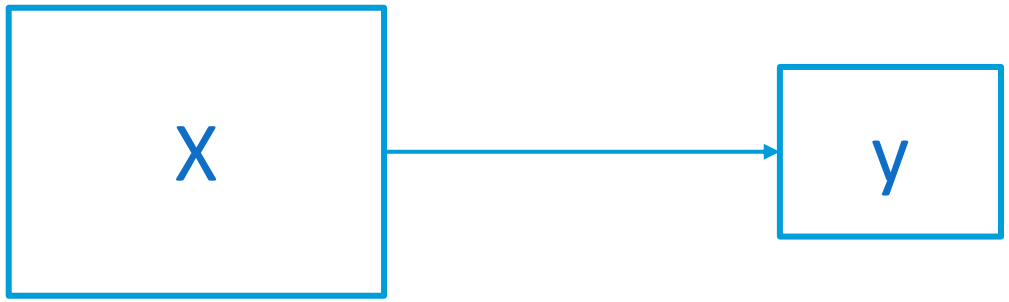
Spam / Bukan Spam

Logistic Regression

Pengambilan Kesimpulan dan Interpretasi vs. Prediksi

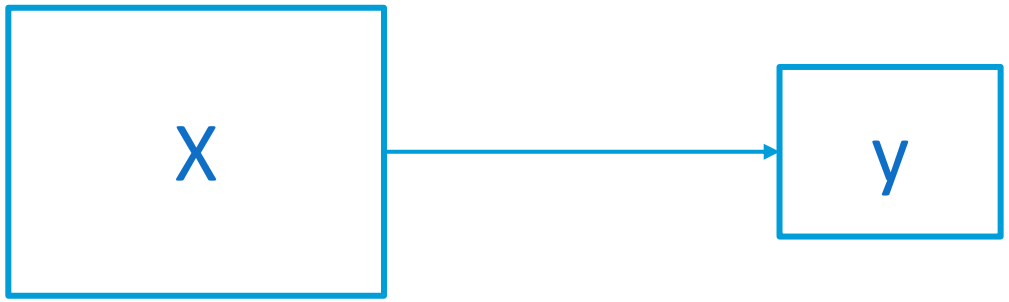
Alur Kerja

Analisis Kelulusan Mahasiswa



- Kumpulkan data
- Cek asumsi
- Jalankan model
- Lihat hasil estimasi parameter
- Lihat variabel yang signifikan
- Pilih variabel, jalankan ulang model
- Evaluasi model dengan: R^2, AIC, BIC
- Setelah model dipilih,
  buat analisis interpretasi dari model.
  Variabel mana yang mempengaruhi kelulusan,
  dan seberapa besar pengaruhnya

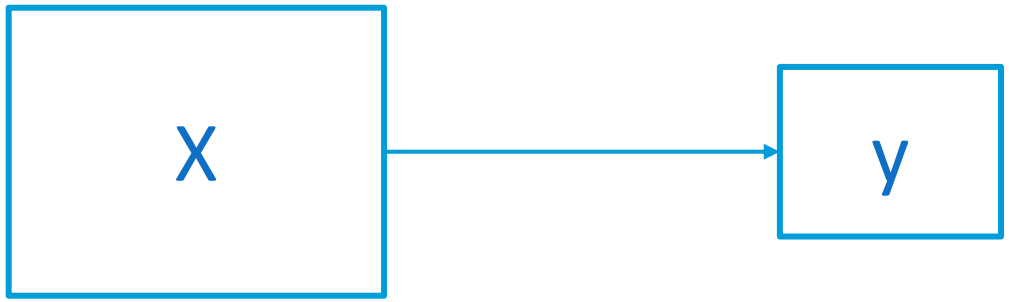
Prediksi Email Spam



- Kumpulkan data
- Transformasi data ke numeric
- Jalankan model
- Tidak begitu peduli dengan nilai estimasi parameter, fokus ke prediksi.
- Tambahkan regularisasi (L1/L2) ke model
- Evaluasi model dengan *cross validation*
- Setelah model dipilih, deploy model ke sistem online untuk memprediksi apakah email baru spam atau tidak.

Arah Pengembangan Lebih Lanjut

Analisis Kelulusan Mahasiswa



Fokus: mendapatkan kesimpulan dan interpretasi yang lebih pas

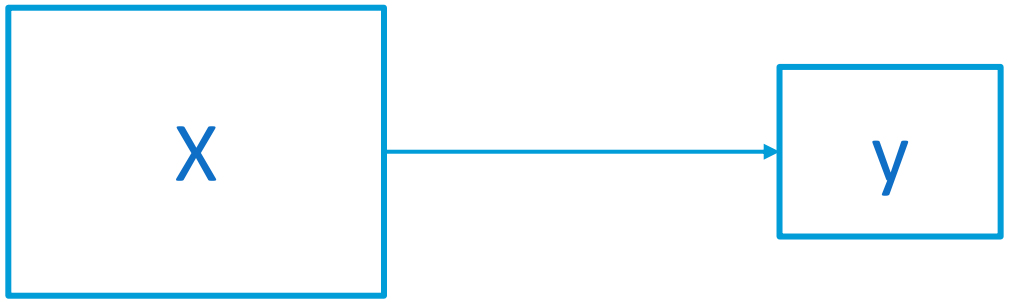
Model-model lebih lanjut:

- Analisis Varians (ANOVA)
- Model berdasarkan distribusi lain, misal dengan: Generalized Linear Model (GLM)
- Analisis Jalur (Path Analisis)
- Analisis Causality

Model-model cenderung linear.

- Mudah di interpretasi

Prediksi Email Spam



Fokus: mendapatkan prediksi yang lebih akurat

Model-model lebih lanjut:

- Tambah variabel, ribuan variabel tidak masalah (contoh: *bag-of-words*)
- Tambahkan non-linearitas ke model
  - SVM dengan kernel model
  - Neural Networks
  - Random Forest

Model kompleks tidak masalah

(Statistika) ∧ (Machine Learning)

Contoh area-area sama-sama di dalami oleh peneliti statistika dan machine learning

Teori Peluang
- Distribusi statistik
- Random Variable
- Expected Values & Variance
- Teorema Bayes.

Estimasi Model
- Maximum Likelihood Estimation
- Consistency, Sufficient Statistics
- Gradient Descent, Newton Methods
- Expectation-Maximization (EM)

Model Linear
- Linear Regression
- Logistic Regression
- Ridge & Lasso Regression
- Variable Reduction, seperti Principal Component Analysis (PCA)
- Generalized Linear Model (GLM)
- Model linear Support Vector Machine

Model Bayesian
- Bayesian Linear/Logistic Regression
- Bayesian GLM
- Model Bayesian Non-Parametrik: Gaussian Processes, Latent Dirichlet Allocation
- Distribution Sampling
- MCMC, Gibbs Sampling

Area Lain
- Probability Density Estimation
- Analisis Cluster

(Statistika) ∧ ¬(Machine Learning)

Contoh area penting bagi peneliti statistika yang kurang didalami peneliti machine learning

Sampling (dari populasi)

- Statistika: penting untuk desain eksperimen atau survei
- ML: data yang dikumpulkan biasanya dari data transaksional. Tidak perlu mengambil sampel

Uji hipotesis

- Statistika: penting untuk menguji dugaan awal dan membantu pengambilan keputusan
- ML: tidak banyak didalami, karena fokus ke prediksi

Analisis varians (ANOVA/MANOVA)

- Statistika: penting untuk menganalisis sumber variasi dari data
- ML: kurang didalami

Model Linear Lanjut

- Statistika: di beberapa aplikasi, perlu menganalisis lebih lanjut hubungan antar variabel. Contoh: analisis jalur (path), analisis survival, model linear dengan asumsi lain (e.g. two stages least square).
- ML: model non-linear biasa dipakai untuk model lebih lanjut

¬(Statistika) ∧ (Machine Learning)

Contoh area penting bagi peneliti machine learning yang kurang didalami peneliti statistika

Kernel Trick

- ML: proyeksi variable ke dimensi lebih tinggi secara lebih efisien. Salah satu cara untuk mendapatkan model non-linear.
- Statistika: tidak banyak didalami, model jadi tidak mudah untuk di-intpretasikan

Neural Networks dan Deep Learning

- ML: model non-linear dengan menumpuk model linear + fungsi non-linear. Populer untuk data yang kompleks (gambar, teks). Bisa *training* representasi dari *raw data*.
- Statistika: Model terlalu kompleks. Susah untuk di-interpretasi.

Inferense Semi-Supervised

- ML: Penting untuk kasus di mana hanya sebagian kecil dari dataset ada label nya, tapi ingin menggunakan data tanpa label juga untuk menambah akurasi dari prediksi. Contoh: prediksi *review* palsu *di e-commerce.*
- Statistika: Tidak banyak dipelajari. Fokus bukan ke prediksi tapi analisis hubungan antar variabel.

Cara Berpikir Matematis vs. Algoritmis

Statistika: Matematis

Peneliti: Dept. of Statistics
Dept. of Mathematics

- Formulasi matematis
- Derivasi rumus
- Pembuktian rumus

Contoh: Linear Regression
Formulasi:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \ \ldots + \beta_d X_{i,d} + \ \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Estimasi parameter: Least Square Estimation

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Matrix Inverse:
Runtime: $O(n^3)$

Machine Learning: Algoritmis

Peneliti: Dept. of Computer Science
Dept. of Statistics

- Formulasi matematis
- Derivasi rumus
- **Analisis *runtime* dari algoritma**

Contoh: Support Vector Machine, Kernel Tricks
Formulasi & Estimasi parameter:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \underbrace{\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(x_i, x_j)}_{\text{Runtime: } O(n^2)}$$

Kalau *n* besar, misal ≥ 100,000. Tidak efisien.
Kernel approximation dengan random features
Runtime: $O(n)$

Kultur Jurnal vs. Konferensi

Statistika: Publikasi di Jurnal

- Annals of Statistics
- Biometrika
- Journal of the American Statistical Association

….

THE ANNALS *of* STATISTICS

Journal of the American Statistical Association

BIOMETRIKA

\+ Lebih *detail*

\- Kurang cepat

ML: Publikasi di Konferensi

- Neural Information Processing System
- International Conference on Machine Learning
- International Conference on Learning Representation

….

NEURAL INFORMATION PROCESSING SYSTEMS

ICML International Conference On Machine Learning

ICLR

\+ Lebih cepat

\- Kurang *detail*

Bahasa Pemrograman

Statistika

Machine Learning

Kesimpulan

Data

Statistika

Pengambilan Kesimpulan
dan Interpretasi

Machine Learning

Prediksi

Hypothesis Testing
ARIMA/ARCH
Linear Models
Support Vector Machine
Semi-supervised Learning
Generalized Linear Model
Maximum Likelihood
Graphical Model
Neural Networks
Random Forest
Population Sampling
ANOVA/MANOVA
Expectation Maximization
Deep Learning
XGBoost
Survival Analysis
Causal Inference
Logistic Regression
Kernel Tricks
Markov Model
Path Analysis
Bootstrapping
Bayesian Model
Reinforcement Learning
Latent Dirichlet Allocation

Referensi

- Blei, David M, Ng, Andrew Y, and Jordan, Michael I. *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3(Jan):993–1022, 2003.
- Breiman, Leo et al. *Statistical modeling: The two cultures (with comments and a rejoinder by the author)*. Statistical science, 16(3):199–231, 2001.
- Friedman, Jerome, Hastie, Trevor, and Tibshirani, Robert. *The elements of statistical learning, volume 1*. Springer series in statistics Springer, Berlin, 2001.
- Harrell, Frank. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.
- Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.
- Wasserman, Larry. *Rise of the machines*.
- Wasserman, Larry. *All of statistics: a concise course in statistical inference.* Springer Science & Business Media, 2013.

Terima Kasih

https://rizal.fathony.com/assets/pdf/statistika_dan_machine_learning.pdf