Machine Learning

Yaya Wihardi, S.Kom., M.Kom.

Email: yayawihardi@upi.edu

Department of Computer Science Education

Universitas Pendidikan Indonesia

Outline

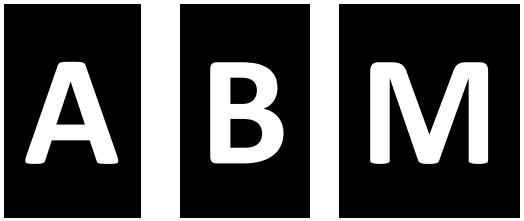
- Apa itu *Machine Learning* (ML)?
- Kenapa Harus Belajar ML?
- Aplikasi ML
- Fase dalam ML
- Permasalahan yang Sering Muncul dalam ML
- Validasi Model
- Perangkat Lunak *ML*

Apa Itu Machine Learning?

- Salah satu disiplin ilmu dalam ilmu komputer
- Tujuan: mempelajari bagaimana membuat sistem komputer yang dapat beradaptasi dan belajar dari pengalaman (Tom Dietterich)
- Pengalaman = Data
- Konsep Pendukung: Probabilistik & Algoritma
- ML berkaitan erat dengan ketidak-pastian

Data

- Contoh data citra karakter:



- Atribut: Karakteristik yang mencerminkan data, misal: tinggi, lebar, jml *cycle*
- Fitur: atribut yang digunakan untuk proses pembelajaran

Label Data	lebar	tinggi	Jml cycle
A	2	3	1
B	2	3	2
M	3	3	0

- Untuk membedakan karakter2 tsb cukup menggunakan fitur: lebar dan jml-cycle

Model/Metode dalam ML

- Regression:
  Least Squares, Logistic Regression
- Instance-based Methods:
  k-Nearest Neighbor (kNN), Learning Vector Quantization (LVQ), Self-Organizing Map (SOM)
- Decision Tree:
  ID3, C4.5, Random Forest
- Bayesian:
  Naive Bayes, Bayesian Belief Network (BBN)
- Kernel Methods:
  Support Vector Machines (SVM), Radial Basis Function (RBF), Linear Discriminant Analysis (LDA)

Model/Metode dalam ML

- Clustering Methods:
  k-Means, Expectation Maximization (EM)
- Association Rule Learning:
  Apriori algorithm
- Artificial Neural Networks:
  Perceptron, Back-Propagation, Self-Organizing Map (SOM), Learning Vector Quantization (LVQ)
- Deep Learning:
  Restricted Boltzmann Machine (RBM), Deep Belief Networks(DBN), Convolutional Network, Stacked Auto-encoders
- Ensemble Methods:
  Boosting, AdaBoost, Random Forest

Jenis Metode Learning

- Supervised Learning
- Unsupervised Learning
- Reinforcement Learning

Jenis Metode Learning

Supervised learning

Agent belajar fungsi yang memetakan input ke output

- Pada tahap training, learning algorithm menerima sekumpulan input(x) beserta output (y) yang diharapkan.
- Sample ini dipakai untuk estimasi fungsinya (f).
- Data yang digunakan harus dilabeli
- Contoh Algoritma: SVM, Logistic Regression, Random Forest, ANN
- Banyak digunakan utk klasifikasi

$$f: x \rightarrow y$$

Contoh Data untuk Supervised Learning

INPUT

Output yg Diharapkan (Class/Label)

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

Jenis Metode Learning

Unsupervised learning

Sebuah learning algorithm menerima sekumpulan data, dan harus menemukan pola-pola di dalamanya.

Misalnya:
Sebuah agent taxi menerima data mengenai laju lalin sepanjang hari. Mungkin ia bisa belajar periode “morning rush hour", “evening rush hour“

- ✓ *Data tidak berlabel*
- ✓ *Contoh Algoritma: PCA, K-Mean, Autoencoders, Self Organizing Maps (SOM), Adaptive Resonance Theory*
- ✓ *Banyak digunakan untuk reduksi dimensi/ekstraksi fitur*

Contoh Data untuk Unsepervised Learning

INPUT

Class Label?

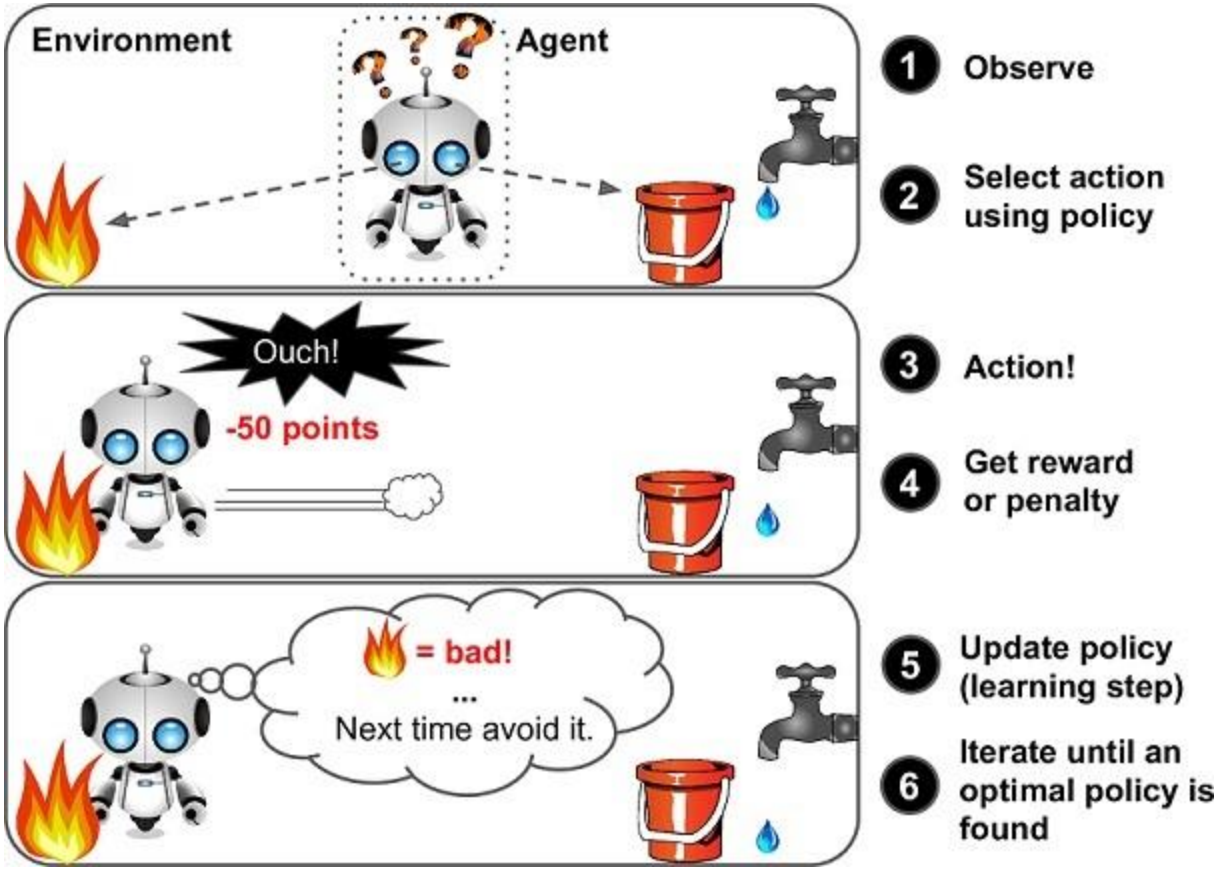
ID	Age	Has_Job	Own_House	Credit_Rating
1	young	false	false	fair
2	young	false	false	good
3	young	true	false	good
4	young	true	true	fair
5	young	false	false	fair
6	middle	false	false	fair
7	middle	false	false	good
8	middle	true	true	good
9	middle	false	true	excellent
10	middle	false	true	excellent
11	old	false	true	excellent
12	old	false	true	good
13	old	true	false	good
14	old	true	false	excellent
15	old	false	false	fair

Jenis Metode Learning

Reinforcement learning

- Sebuah agent menerima input data dan harus mengambil tindakan.
- Agent lalu menerima reinforcement signal (mis. good, bad) sebagai akibat tindakan.
- Learning algorithm memodifikasi agent function untuk memaksimalkan signal “good".

Reinforcement Learning



Gabungan?

- Semi-Supervised Learning
  - Supervised+Unsupervised
  - Banyak digunakan utk melabeli data supervised yg terbatas

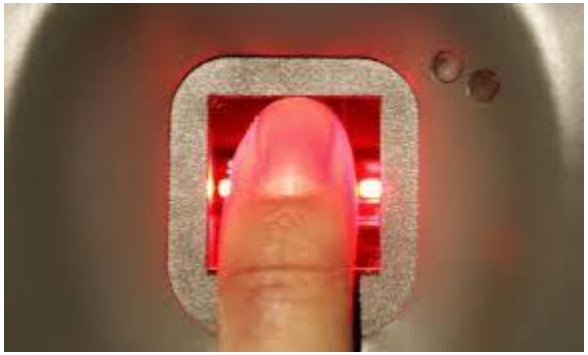
Contoh Data pada Semi-supervised Learning

INPUT

Class/Label

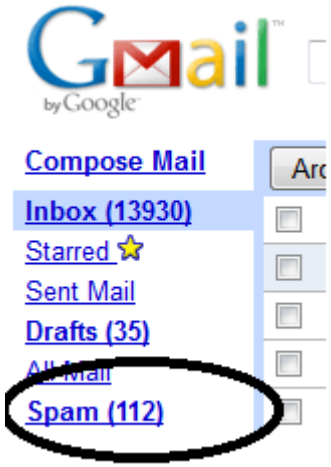
ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	good	?
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	?
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	?
11	old	false	true	excellent	?
12	old	false	true	good	?
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

Contoh Aplikasi Machine Learning



Pengenalan Sidik Jari

Deteksi Area Wajah



Klasifikasi Email Spam

Pencarian Asosiasi Antar Produk yang Dibeli Pelanggan

Kenapa Harus Belajar ML?

- Era komputerisasi sudah (hampir) usai, hampir semua data sudah tersedia dalam bentuk data digital, bahkan sekarang kita sudah berada di era globalisasi informasi.
- Proses produksi data semakin cepat: transaksi keuangan, tweet, foto, video, depth-image, dll.
- Data semakin mudah didapat dan jumlahnya terus bertambah.
- Apa yang bisa dilakukan dengan data tsb? Melalui ML, kita dapat membangun sistem yang cerdas yang mampu mengolah dan menganalisis data tersebut menjadi informasi yang sangat berharga, seperti: mengenali pola, prediksi, klasifikasi, *clustering*, segmentasi, dll.

Aplikasi ML

Kapan ML Dibutuhkan?

- Ketika Keahlian manusia (hampir) tidak mungkin digunakan: Navigasi di Mars
- Manusia tidak mampu menjelaskan keahliannya secara pasti: Pengenalan Suara, Pengenalan Wajah
- Solusi dari masalah selalu berubah dari waktu ke waktu: Routing Network
- Solusi masalah perlu disesuaikan dengan kasus-kasus tertentu: Biometric

Bidang Aplikasi ML

- **Learning associations**
  Contoh: Menemukan asosiasi antar produk yang dibeli pelanggan
- **Classification**
  Contoh: Klasifikasi resiko pengguna dalam pengajuan kartu kredit ke dalam low dan high
- **Regression**
  Contoh: Prediksi harga saham enam bulan ke depan
- **Unsupervised learning**
  Contoh: Segmentasi citra satelit berdasarkan area tutupannya (land cover)
- **Reinforcement learning**
  Contoh: Setelah beberapa kali dijalankan, sebuah game dapat mempelajari langkah-langkah untuk menang

Fase dalam ML

- Pra-pengolahan (pre-processing)
  Proses ekstraksi dan pemilihan fitur dari atribut data
- Pelatihan (Training): Supervised, Unsupervised, Reinforcement
  Proses pembangunan model dengan menggunakan data latih.
- Pengujian (Testing)
  Proses pengujian model dengan data uji.
- Evaluasi: cross-validation, bootstrapping, dll
  Proses evaluasi kehandalan model yang dibangun, biasanya digunakan metric evaluasi khusus, dll.

Proses Training Pada Proses Klasifikasi

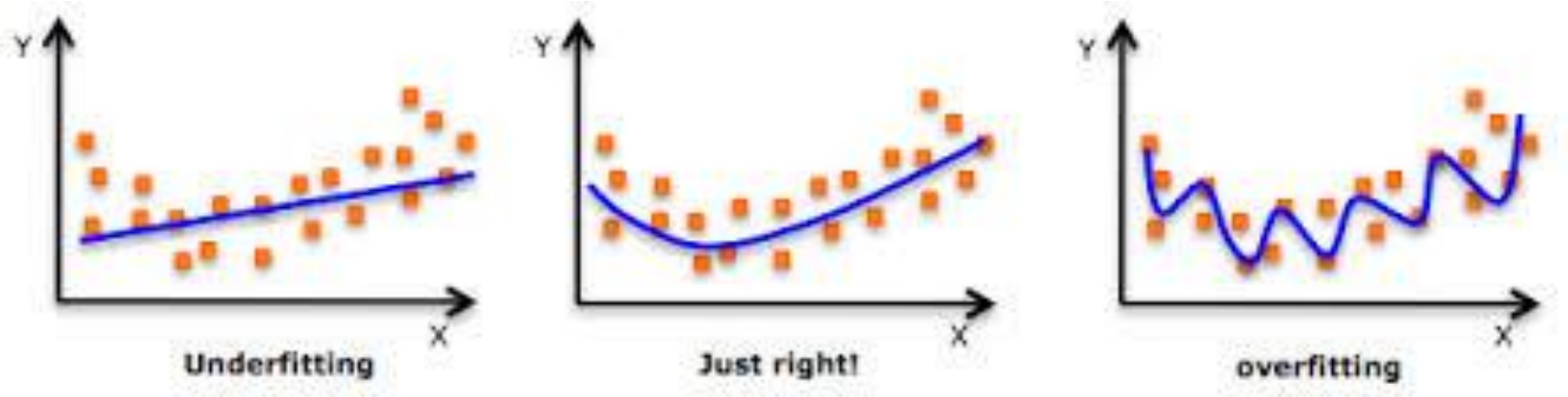
- Setiap model memiliki parameter yang perlu di-tuning
- Jika mesin classifier dianalogikan sebagai mesin, proses tuning parameter adalah proses mencari settingan yang tepat sedemikian hingga mesin tsb dapat melakukan klasifikasi sesuai data latih yang diberikan
- Proses Training biasanya memakan waktu yang lama, bisa ber jam-jam, bahkan berhari-hari
- Biasanya menggunakan algoritma pembelajaran tertentu, salah satunya yaitu: Gradient Descent
- Pada setiap siklus/iterasi proses training, model sesegera mungkin divalidasi dengan menggunakan data validasi, guna mengetahui seberapa bagus hasil pembelajaran sampai pada iterasi tersebut
- Setelah proses training selesai, nilai parameter hasil learning disimpan dalam file utk digunakan pada tahap produksi

Testing/Evaluasi pada Proses Klasifikasi

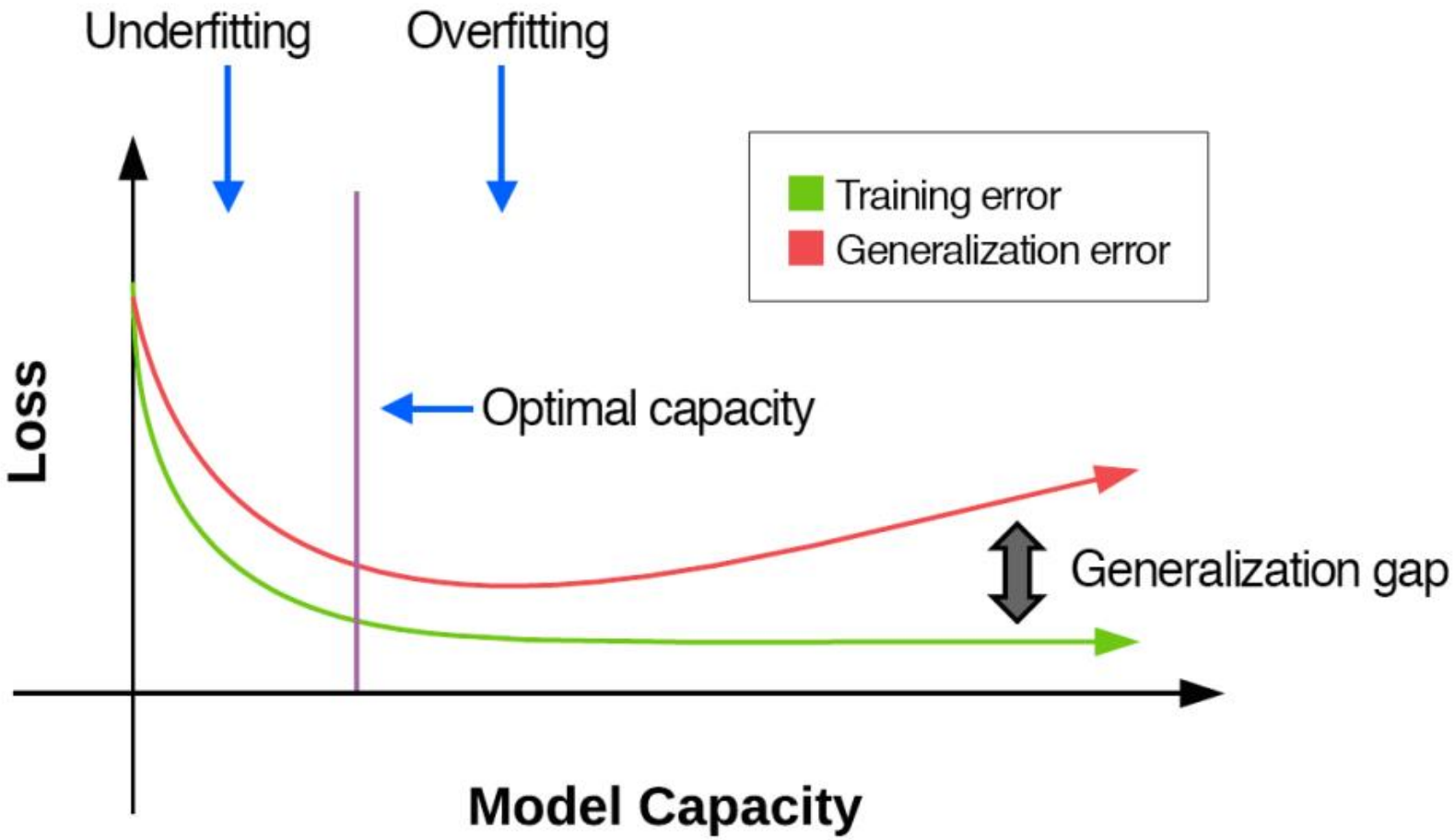
- Setelah proses training selesai, model di uji dengan menggunakan testing set
- Output dari mesin/model dibandingkan dengan groundtruth/label yang sesungguhnya dari testing set
- Setiap output yang berbeda dengan labelnya, maka dihitung sebagai sebuah error
- Data hasil pengujian tersebut kemudian dinyatakan dalam perhitungan accuration, precision, recal, dan atau f-measure yang menyatakan seberapa bagus model yg dibangun

Permasalahan yang Sering Muncul

- Underfitting
  Solusi: Memperbaiki proses learning, Meningkatkan kompleksitas model
- Overfitting
  Solusi: Regularisasi, Menambah jumlah data, menurunkan kompleksitas model



Analisis Hasil Pengujian pada Proses Klasifikasi

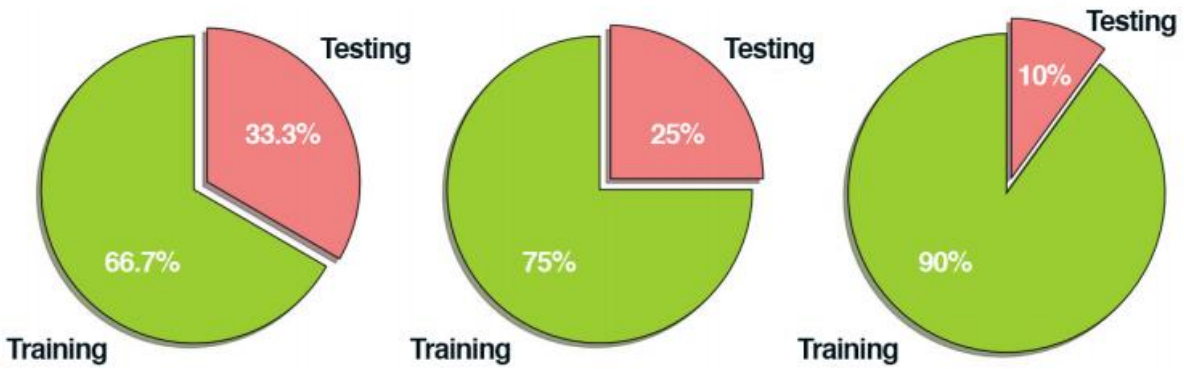


Pembangunan Dataset

- Kategori/Kelas harus terdefinisi dengan jelas,
  Misal: dog, cat, panda
- Jumlah data per kategori harus uniform, atau relatif sama
- Jika jumlah data setiap kelas berbeda, maka terjadi imbalance, harus dilakukan penanganan khusus, namun sebaiknya hindari imbalance ini
- Contoh Dataset:
  - MNIST
  - CIFAR-10
  - SMILES
  - CALTECH-101
  - ImageNET, dll

Pembagian Dataset

- Dataset Dibagi menjadi 2 bagian:
  - Training Set: Utk training/pembangunan model
  - Test Set: Utk evaluasi model hasil training
- **Important!** Trainig Set dan Test Set, masing-masing independen, tidak ada overlapping data, artinya benar-benar terpisah
- Rasio pembagian yang biasa dilakukan:



- Mengingat Test Set hanya digunakan sebagai test, padahal dalam proses training dibutuhkan proses validasi, maka biasanya Training Set dibagi menjadi 2 bagian lagi, yaitu training set dan validation set

Validasi Model

Tujuan:

- Menentukan teknik *sampling* data
- Memvalidasi dan mengukur tingkat kehandalan model yang dibangun

Metode:

- K-Fold Cross-Validation
- Bootstraping

K-Fold Cross Validation

- Data dibagi menjadi K partisi, misal K=5

- Eksperimen dilakukan sebanyak K kali

E1→ Data testing: P1 Data Latih: P2,P3,P4,P5
E2→ Data testing: P2 Data Latih: P1,P3,P4,P5
E3→ Data testing: P3 Data Latih: P1,P2,P4,P5
E4→ Data testing: P4 Data Latih: P1,P2,P3,P5
E5→ Data testing: P5 Data Latih: P1,P2,P3,P4

- Akurasi maupun presisi merupakaan rataan dari K kali eksperimen
  Akurasi = (AE1+AE2+AE3+AE4+AE5)/5
- Biasanya digunakan utk data yang banyak

Bootstraping

- Eksperimen dilakukan sebanyak K kali
- Data dibagi dua secara random dari himpunan semesta pada setiap kali eksperimen dengan rasio tertentu, misal 30% data test, 70% data latih
- Sebuah data mungkin saja menjadi data latih ataupun testing beberapa kali pada eksperimen yang berbeda
- Akurasi maupun presisi merupakan rataan dari akurasi setiap eksperimen

ML Software

- Matlab
- R
- Octave
- Shogun
- Orange
- Weka
- Rapidminer

Occam Razor

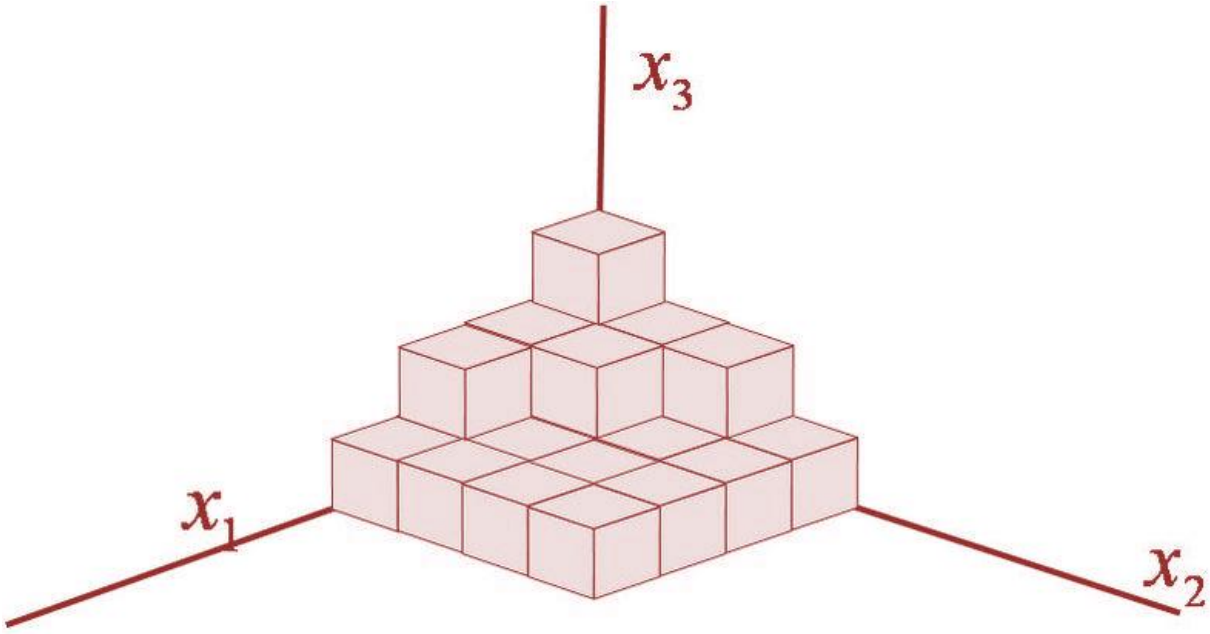
Occam's Razor

"plurality must never be posited without necessity"

William of Ockham
1288 – 1348

Curse of Dimensionality

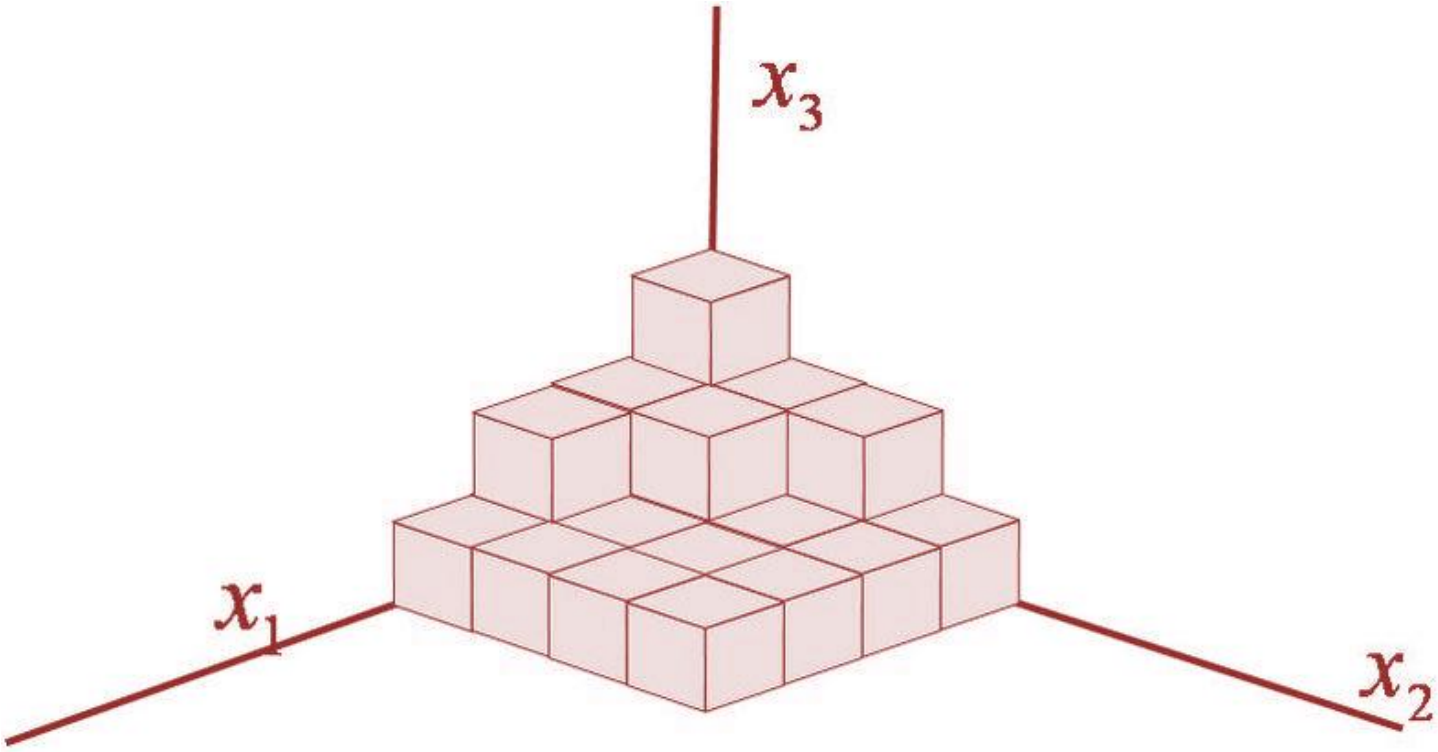
- Fenomena dimana jumlah atribut terus bertambah, sedangkan jumlah sampel data terbatas, maka akurasi klasifikasi ini pada titik tertentu akan menurun.
- Fenomena pertama kali dibahas oleh Bellman di papernya th.1961.



Curse of Dimensionality

- Tingginya dimensi data membuat masalah jadi sulit, Mengapa ?
- Observasi meteorologi untuk menentukan muncul tidaknya kabut berkisar 26 atribut
- Data biomedis yang dipakai untuk memprediksi efektifitas terapi interferon pada pasien hepatitis C kronis berkisar 30.
- Dimensi input data hasil feature extraction pada tulisan tangan adakalanya lebih dari 700 (feature extraction memakai Peripheral Distributed Contributivity, menghasilkan data berdimensi 768 dari mesh 64×64).
- Bahkan ada diantaranya yang berdimensi ribuan, seperti data microarray, karena gen manusia berkisar 22 ribu

Curse of Dimensionality



Curse of Dimensionality

- Jika dimensi data D, dan tiap sumbu variabel dibagi ke dalam M interval, diperlukan setidaknya M^D data pada training set yang mengisi tiap sel, agar pemetaan non linear itu bisa berjalan dengan baik.
- Konsekuensi: jika dimensi data itu ditingkatkan menjadi D+1, maka banyaknya data yang diperlukan akan menjadi M^D * M = M^(D+1) ,
- Data yang diperlukan meningkat secara eksponensial seiring dengan meningkatnya dimensi data.
- Pada umumnya data yang tersedia dan dapat dipakai sebagai training set berjumlah terbatas.
- Hal ini mengakibatkan, jika tidak semua sel pada ruang vektor itu terisi oleh data, dengan kata lain, keberadaan data sangat sparse yg menyebabkan model pemetaan yang dibangun tidak akan bekerja dengan optimal atau baik.

Thank You