MACHINE LEARNING

Materi#3 : Data Analytics

Deskripsi

Tujuan dari pembelajaran Materi ke-3 adalah; Mahasiswa akan mendapatkan pengetahuan dasar analisis data yang digunakan pada Machine Learning. Mahasiswa akan mendapatkan pemahaman terhadap tahapan umum untuk analisis data (CPMK1)

Capaian Pembelajaran

Pada pertemuan ini, kita akan mempelajari:

- Pengenalan Analisis Data
- Nilai, Atribut dan Transformasi
- Ruang Konsep
- Linear Separability
- Seleksi Fitur
- Classification, Association dan Clustering
- Mengukur Kinerja
- Evaluasi Model
- Kategori Jenis Algoritma
- Tahapan Analisis

Referensi

- Jan Wira Gotama Putra, Pengenalan Konsep Pembelajaran Mesin dan Deep Learning, Tokyo, Jepang: -, 2020.

Agenda

- Memahami proses dan tahapan analisis data

Pengenalan Data Analitik

- **Mengenali Masalah**. Masalah; ketika tujuan yang diinginkan tidak tercapai (*current state* bukanlan *desire state*).
- **Problem Solving.** Agar permasalahan bisa diselesaikan (*current state* menjadi *desire state*).
- **Pahami Domain Masalah.** Tiap bidang (domain) punya masalah yang berbeda. Menggunakan teknik *machine learning* tanpa tahu domain aplikasi = buta. Contoh; NLP → voice to teks, klasifikasi teks. Komputasi satu domain dengan lainnya pasti berbeda.
- **Machine Learning adalah inferensi berdasarkan data**; raw data / data mentah adalah sekumpulan fakta yang tidak arti sama sekali; tidak rapi, missing value, tidak ada label.
- **Preprocessing**; merapikan data (dataset), mencari fitur yang sesuai
- **Membangun Model;** menggunakan algoritma untuk mengetahui pola
- **Performance Measure;** seberapa 'bagus' model yang kita bangun (dalam tabel); apakah sesuai dengan pola yang ada di data, sesuai tujuan analisis (dalam bentuk kurva, grafik); Apakah ada outlier, bagaimana pengaruhnya

Kesimpulan → PAHAMI DOMAIN PERMASALAHAN!!

Nilai Atribut dan Transformasi

Tipe Atribut :

1. **Nominal**. Nilai atribut bertipe nominal tersusun atas simbol-simbol yang berbeda, yaitu suatu himpunan terbatas. Sebagai contoh, fitur *outlook* pada Tabel 3.1 memiliki tipe data **nominal** yaitu nilainya tersusun oleh himpunan {*sunny, overcast, rainy*}. Pada tipe nominal, tidak ada urutan ataupun jarak antar atribut. Tipe ini sering juga disebut **kategorial** atau **enumerasi**. Secara umum, tipe *output* pada *supervised learning* adalah data nominal.

2. **Ordinal**. Nilai ordinal memiliki urutan, sebagai contoh $4 > 2 > 1$. Tetapi jarak antar suatu tipe dan nilai lainnya tidak harus selalu sama, seperti $4-2 \neq 2-1$. Atribut ordinal kadang disebut sebagai **numerik** atau **kontinu**.

3. **Interval**. Tipe interval memiliki urutan dan *range* nilai yang sama. Sebagai contoh $1-5, 6-10, dst$. Kita dapat mentransformasikan/ mengkonversi nilai numerik menjadi nominal dengan cara merubahnya menjadi interval terlebih dahulu. Lalu, kita dapat memberikan nama (simbol) untuk masing-masing interval. Misalkan nilai numerik dengan *range* $1-100$ dibagi menjadi 5 kategori dengan masing-masing interval adalah $\{1-20, 21-40, \ldots, 81-100\}$. Setiap interval kita beri nama, misal interval $81-100$ diberi nama *nilai* A, interval $61-80$ diberi nama *nilai* B.

4. **Ratio**. Tipe *ratio* (rasio) didefinisikan sebagai perbandingan antara suatu nilai dengan nilai lainnya, misalkan massa jenis (fisika). Pada tipe *ratio* terdapat *absolute zero* (semacam *ground truth*) yang menjadi acuan, dan *absolute zero* ini memiliki makna tertentu.

Label

Tabel 3.1

Fitur

id	outlook	temperature	humidity	windy	play (class)
1	sunny	hot	high	false	no
2	sunny	hot	high	true	no
3	overcast	hot	high	false	yes
4	rainy	mild	high	false	yes
5	rainy	cool	normal	false	yes
6	rainy	cool	normal	true	no
7	overcast	cool	normal	true	yes
8	sunny	mild	high	false	no
9	sunny	cool	normal	false	yes
10	rainy	mild	normal	false	yes
11	sunny	mild	normal	true	yes
12	overcast	mild	high	true	yes
13	overcast	hot	normal	false	yes
14	rainy	mild	high	true	no

Fitur Vektor

Representasi observasi data

Dataset

Atribut

Ruang Konsep

Dengan data yang diberikan, kita ingin melakukan generalisasi aturan/ konsep yang sesuai dengan data. Hal ini disebut sebagai *inductive learning*. Cara paling sederhana untuk *inductive learning* adalah mengenumerasi seluruh kemungkinan kombinasi nilai sebagai *rule*, kemudian mengeleminasi *rule* yang tidak cocok dengan contoh. Metode ini disebut *list-then-eleminate*. Silahkan baca buku Tom Mitchell [4] untuk penjelasakn lebih rinci. Kemungkinan kombinasi nilai ini disebut sebagai ruang konsep (***concept space***). Sebagai contoh pada Tabel 3.1 himpunan nilai masing-masing atribut yaitu:

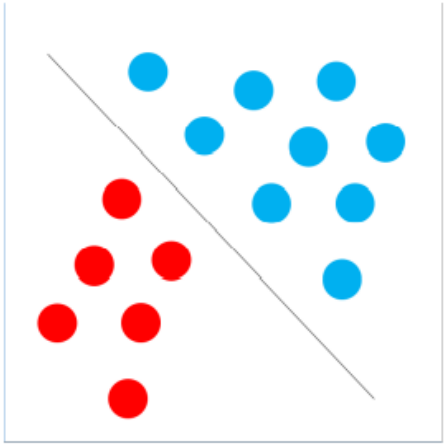
- $outlook = \{sunny, overcast, rainy\}$
- $temperature = \{hot, mild, cold\}$
- $humidity = \{high, normal\}$
- $windy = \{true, false\}$
- $play = \{yes, no\}$

sehingga terdapat $3 \times 3 \times 2 \times 2 \times 2 = 72$ kemungkinan kombinasi. Tentunya kita tidak mungkin mengenumerasi seluruh kemungkinan kombinasi nilai karena secara praktikal, atribut yang digunakan banyak. Terlebih lagi, apabila mengenumerasi kombinasi atribut bertipe numerik.

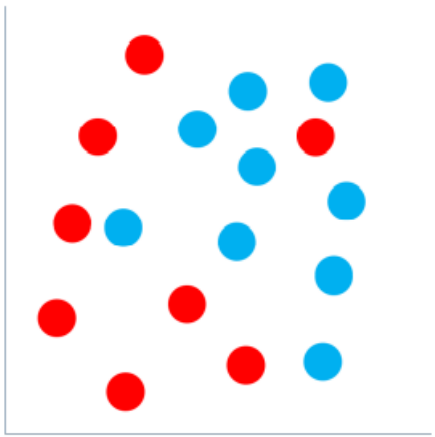
Linear Separability

id	humidity	windy	swim (class)
1	high	high	yes
2	normal	normal	no

?



Linearly Separable



Non Linearly Separable

- Data yang non linearly separable harus ditransformasikan menjadi linear separable
- Teknik transformasi yang umumnya digunakan kernel function ; radial basis function
- Teknik ini dilakukan dengan menambah fitur (2 fitur jadi 3 fitur)

Seleksi Fitur

Menyederhanakan fitur terkadang diperlukan, dengan alasan:

- Menyederhanakan data atau model agar mudah dianalisis
- Mengurangi waktu training (mengurangi kompleksitas model dan inferensi)
- Menghindari curse of dimensionality
- Menghapus fitur yang tidak informatif
- Meningkatkan generalisasi dengan mengurangi overfitting

Contoh; menghapus atribut yang memiliki nilai 0, karena fitur ini tidak memiliki nilai informatif

Classification, Association dan Clustering

- Pada supervised learning, prediksi kelas berdasarkan feature vector
- Setiap feature vector berkoresponden dengan kelas tertentu
- Proses mencari kelas yang berkorespondensi terhadap suatu input disebut **klasifikasi (classification)**

  Contoh; Kategori buah

- Mencari hubungan satu atribut dengan atribut yang lain disebut **Asosiasi (association)**

  Outlook = sunny maka sebagian besar humidity = high

- Pada **clustering / clustering** tidak ada kelas yang berkorespondensi

  Contoh; Mengelompokkan barang di supermarket

Mengukur Kinerja

- Mengukur kinerja dengan kuantitatif; optimalisasi utility function pada saat dilatih; minimalisasi nilai error (semua tergantung algoritma)
- Validasi dan testing data → evaluasi prediksi final dengan performance measure (menggunakan metrik; akurasi, presisi, recall, F1 Measure
- Semua tergantung domain masalah.

  Berikut contoh akurasi untuk klasifikasi

$$\text{akurasi} = \frac{\#input \text{ diklasifikasikan dengan benar}}{\text{banyaknya data}}$$

Evaluasi Model

- **Data splitting**
- **Overfitting**; keadaan ketika model memiliki kinerja baik hanya untuk training data tetapi tidak memiliki kinerja baik untuk unseen example **dan Underfitting;** keadaan ketika model memiliki kinerja buruk baik untuk training data dan unseen example
- **Cross validation;** apakah model memiliki generalisasi yang baik(mampu memiliki kinerja baik pada unseen examples)

Kategori Jenis Algoritma

Algoritma pembelajaran mesin dapat dibagi menjadi beberapa kategori. Dari sudut pandang apakah algoritma memiliki parameter yang harus dioptimasi, dapat dibagi menjadi:[8]

1. Parametrik. Pada kelompok ini, kita mereduksi permasalahan sebagai optimisasi parameter. Kita mengasumsikan permasalahan dapat dilambangkan oleh fungsi dengan bentuk tertentu (e.g., linier, polinomial, dsb). Contoh kelompok ini adalah model linier.
2. Non parametrik. Pada kelompok ini, kita tidak mengasumsikan permasalahan dapat dilambangkan oleh fungsi dengan bentuk tertentu. Contoh kelompok ini adalah *Naive Bayes*, *decision tree* (ID3) dan *K-Nearest Neighbors*.

Dari sudut pandang lainnya, jenis algoritma dapat dibagi menjadi:

1. Model linear, contoh regresi linear, regresi logistik, *support vector machine*.
2. Model probabilistik, contoh *Naive Bayes*, *hidden markov model*.
3. Model non-linear, yaitu (*typically*) *artificial neural network*.

Tahapan Analisis

1. Memutuskan tujuan analisis data (*defining goal*)
2. Mendapatkan data
3. Merapihkan data
4. Merepresentasikan data sebagai *feature vector*
5. Melakukan transformasi dan/atau *feature selection* (mengurasi dimensi *feature vector*)
6. Melatih model (*training*) dan menganlisis kinerjanya pada *validation* (*development*) *data*.
7. Melakukan *testing* dan analisis model baik secara kuantitatif dan kualitatif
8. Menyajikan data (presentasi)

Tugas#3

3.1. Konversi atribut
Sebutkan dan jelaskan macam-macam cara untuk mengkonversi atribut! Sebagai contoh, numerik-nominal dan nominal-numerik.

3.2. Transformasi data
Sebutkan dan jelaskan macam-macam cara transformasi data (e.g. merubah *non-linearly separable* menjadi *linearly separable*)

3.3. Seleksi fitur
Bacalah algoritma seleksi fitur pada *library* sklearn. Jelaskan alasan (*rationale*) dibalik penggunaan tiap algoritma yang ada!

3.4. Inductive Learning
Jelaskanlah algoritma *list-then-eliminate* dan *candidate-elimination*!

3.5. Tahapan analisis
Agar mampu memahami tahapan analisis data dan pembelajaran mesin secara lebih praktikal, kerjakanlah tutorial berikut!
https://scikit-learn.org/stable/tutorial/basic/tutorial.html