BINUS UNIVERSITY ONLINE LEARNING

People
Innovation
Excellence

LECTURE NOTES

COMP8014

Knowledge Data Discovery

Exploring Data

LEARNING OUTCOMES

L0 1: Mahasiswa diharapkan mampu menjelaskan konsep dasar Knowledge Data Discovery.

LO 2: Mahasiswa diharapkan mampu menggunakan teknik eksplorasi data dan preprocessing.

OUTLINE MATERI :

1. Apakah Data itu?
2. Dataset, objek, atribut
3. Deskripsi data statistik
4. Visualisasi data
5. Kesamaan dan ketidaksamaan data
6. Kualitas data

ISI MATERI

Apakah data itu?

Kumpulan objek data dan atributnya.

- Atribut adalah properti atau karakteristik suatu objek

  Contoh: warna mata seseorang, suhu, dll.
- Atribut juga dikenal sebagai variabel, bidang, karakteristik, atau fitur
- Kumpulan atribut menggambarkan suatu objek
- Objek juga dikenal sebagai record, point, case, sample, entity, atau instance

Tipe dari Atribute?

- Nominal
- Ordinal
- Interval
- Ratio

Type dari Atribut berdasarkan kontinyu atau tidaknya?

- Discrete
- Continuous

Karakteristik Deskriptif di Data Mining?

- Central tendency characteristics:
  - mean,
  - median,
  - mode
- Data dispersion characteristics :
  - quartiles,
  - interquartile range (IQR),
  - variance

Pengukuran Dispersi Data?

- Quartiles, outliers and boxplots
  - Quartiles: Q_1 (25^{th} percentile), Q_3 (75^{th} percentile)
  - Inter-quartile range: IQR = $Q_3 - Q_1$
  - Five number summary: min, Q_1, M, Q_3, max
  - Boxplot: ends of the box are the quartiles, median is marked, whiskers, and plot outlier individually
  - Outlier: usually, a value higher/lower than 1.5 x IQR
- Variance and standard deviation (*sample: s, population: σ)*
  - Variance: (algebraic, scalable computation)
  - Standard deviation *s (or σ)* is the square root of variance s^2 *(or* σ^2*)*

Similarity and Dissimilarity?

- Similarity /Kesamaan
  - Ukuran numerik seberapa mirip dua objek data.
  - Lebih tinggi bila benda lebih mirip.
  - Sering jatuh di kisaran [0,1]
- Dissimilarity/ Perbedaan
  - Ukuran numerik seberapa berbeda dua objek data
  - Jatuh pada saat objek lebih mirip
  - Ketidaksamaan minimal sering 0
  - Batas atas bervariasi
- Proximity/Kedekatan mengacu pada kesamaan atau perbedaan

Jenis Jarak?

- Euclidean distance
- Minkowski Distance

Kualiti Data?

- Masalah kualitas data macam apa?
- Bagaimana kita bisa mendeteksi masalah dengan data?
- Apa yang bisa kita lakukan tentang masalah ini?
- Contoh masalah kualitas data:
    - Kebisingan dan outlier
    - Nilai yang hilang
    - Data duplikat

SIMPULAN

1. Data adalah kumpulan objek dan atributnya.
2. Jenis Atribut: Nominal, Ordinal, Interval dan Rasio.
3. Data dapat diskrit atau kontinyu.
4. Beberapa jenis struktur dataset disajikan.
5. Deskripsi statistik digunakan untuk mengetahui kecenderungan sentral dan dispersi data.
6. Visualisasi data untuk lebih memahami data.
7. Beberapa rumus jarak untuk mengukur kesamaan / ketidaksamaan.
8. Kualitas data mempengaruhi hasil analisis data.

DAFTAR PUSTAKA

1. Han, J., Kamber, M., & Pei, Y. (2006). “Data Mining: Concepts and Technique”. Edisi 3. Morgan Kaufman. San Francisco
2. Tan, P.N., Steinbach, M., & Kumar, V. (2006). “Introduction to Data Mining”. Addison-Wesley. Michigan
3. Witten, I. H., & Frank, E. (2005). “Data Mining : Practical Machine Learning Tools and Techniques”. Second edition. Morgan Kaufmann. San Francisco