Data Mining Tasks
Task 1: Clustering

Dr. Bambang Purnomosidi D. P.
https://zimera-systems.com

Software and Data Technologies

Agenda

zimera systems

1. Pengertian *Clustering* dan *Cluster Analysis*
2. Manfaat dan Penggunaan *Clustering*
3. Teknik dan Algoritma *Clustering*
4. K-Means *Clustering*
5. Algoritma K-Means *Clustering*
6. Penyiapan Data untuk K-Means *Clustering*
7. Implementasi

Clustering dan *Cluster Analysis*

zimera systems

- *Clustering == Clustering Analysis == Data Clustering == Segmentation Analysis == Taxonomy Analysis == Unsupervised Classification.*
- **Clustering** merupakan task untuk mengelompokkan sekumpulan obyek data ke dalam beberapa grup yang disebut sebagai *cluster* berdasarkan kemiripan sehingga tingkat keterhubungan antar data dalam cluster yang sama - kuat, sedangkan dengan cluster yang berbeda - lemah.
- Termasuk dalam kategori *unsupervised learning*:
  - data tanpa label
  - untuk menemukan pola
  - hanya menyediakan variabel input (independent var), tidak ada dependent variable

Clustering dan *Cluster Analysis* (2)

zimera systems

- Merupakan bagian dari EDA (*Exploratory Data Analysis*) di Statistika.
- *Clustering* juga melibatkan visualisasi data sebagai bagian dari EDA.
- *Clustering* biasanya digunakan untuk segmentasi data.
- *Clustering* bermanfaat untuk melihat data dari sisi:
  - **arti penting data**: memperluas pengetahuan dalam suatu domain, misal di dunia kedokteran, bisa digunakan untuk melihat reaksi golongan darah tertentu terhadap virus COVID-19
  - **manfaat data*:*** bermanfaat sebagai suatu perantara untuk proses dalam *data pipeline*, misalnya segmentasi konsumen bisa menunju ke langkah selanjutnya untuk program kampanye iklan yang lebih sesuai dengan target.

Manfaat dan Penggunaan *Clustering*

zimera systems

- *Clustering* digunakan untuk pengenalan pola segmentasi data.
- Semua bidang yang memungkinkan untuk memanfaatkan segmentasi data merupakan bidang yang potensial sebagai pengguna *clustering*:
  - **Marketing**: segmentasi pasar dan konsumen.
  - **Kedokteran**: cluster untuk diagnosis, cluster pengembangan penyakit menular
  - **Edukasi**: mengenali (maha)siswa yang memerlukan perhatian khusus.
  - **Biologi**: taksonomi spesies.
  - **Keuangan**: mendeteksi *fraud* dalam transaksi kartu kredit, mengenali pola cuci uang, dan lain-lain.
  - **IT**: social network analysis - mengetahui sebaran twit dan menganalisis twit yang merupakan bot atau bukan real.

Teknik dan Algoritma *Clustering*

zimera systems

Terdapat beberapa teknik dan algoritma *clustering*:

- Affinity Propagation
- Agglomerative Clustering
- BIRCH
- DBSCAN
- K-Means
- Mini-Batch K-Means
- Mean Shift
- OPTICS
- Spectral Clustering
- Mixture of Gaussians

K-Means *Clustering*

zimera systems

- Merupakan salah satu algoritma clustering
- Bertujuan untuk mempartisi dataset (n hasil observasi) ke dalam k jumlah cluster dan masing-masing hasil observasi menjadi bagian dari cluster yang mempunyai nilai centroid terdekat.
- Dimulai dari ide tahun 1956 (Hugo Steinhaus), algoritma standar dibuat oleh Stuart Lloyd (1957), Edward W. Forgy juga membuat algortima (1965). Istilah **k-means** oleh James Macqueen (1967).
- Sering disebut juga **Lloyd-Forgy Algorithm**.

Algoritma K-Means *Clustering*

zimera systems

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

Sumber: https://realpython.com/k-means-clustering-python/

Penyiapan Data K-Means *Clustering*

zimera systems

- Penyiapan data meliputi data numerik serta kategori (datam bentuk numerik - misal **laki-laki: 0, perempuan: 1**)..
- Setelah itu siapkan dalam format CSV (atau SpreadSheet / Excel).
- Data ini kemudian akan dinormalisasi. Secara manual, perlakuan untuk normalisasi ini cukup rumit, *tedious*, dan *error-prone*. Lihat materi pada demo untuk normalisasi ini.
- Normalisasi hanya diperlukan jika terdapat perbedaan *range*. Contoh, data usia 0-100, data pendapatan per bulan 2 juta - 50 juta. Range ini jauh, sehingga harus dinormalisasi. Jika tidak, bisa mempengaruhi hasil.

Implementasi K-Means *Clustering*

zimera systems

1. Perhitungan manual
2. Menggunakan Python

Demo

Penutup

zimera
systems

Thanks!

Got question(s)?

zimera.systems@gmail.com
https://zimera-systems.com

Credits: This template includes Icons by **Flaticon** and images by **Freepik**