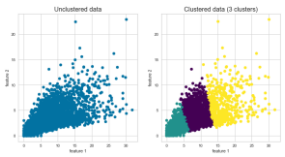
MACHINE LEARNING: UNSUPERVISED LEARNING

10S3001 – Artificial Intelligence

by Samuel I. G. Situmeang



Faculty of Informatics and Electrical Engineering

OBJECTIVES

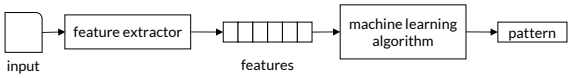
- Students are able to explain the concept of unsupervised learning, including its essential components and applications.
- Students are able to define clustering as an unsupervised learning task and discuss common classification algorithms and their underlying principles.
- Students are able to describe the K-Means process and their use in clustering.
- Students are able to explain various metrics used to evaluate the performance of clustering algorithms.

- Mahasiswa mampu menjelaskan konsep pembelajaran tanpa pengawasan, termasuk komponen-komponen penting dan aplikasinya.
- Mahasiswa mampu mendefinisikan klasterisasi sebagai tugas pembelajaran tanpa pengawasan dan membahas algoritma klasifikasi umum dan prinsip-prinsip yang mendasarinya.
- Mahasiswa mampu menjelaskan proses K-Means dan penggunaannya dalam klasterisasi.
- Mahasiswa mampu menjelaskan berbagai metrik yang digunakan untuk mengevaluasi kinerja algoritma klasterisasi.

UNSUPERVISED LEARNING

WHAT IS UNSUPERVISED LEARNING?

- **Unsupervised learning** finds **hidden patterns** or **intrinsic structures** in **data**.

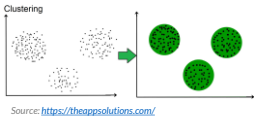


- Compared to supervised learning where training data is labeled with the appropriate classifications, unsupervised learning learns the **relationships** between elements in a dataset without labeling the data as any particular classification.

- Pembelajaran tanpa supervisi menemukan pola tersembunyi atau struktur intrinsik dalam data.
- Dibandingkan dengan pembelajaran dengan supervise di mana data pelatihan diberi label dengan klasifikasi yang sesuai, pembelajaran tanpa supervisi mempelajari hubungan antara elemen dalam kumpulan data tanpa memberi label data sebagai klasifikasi tertentu.

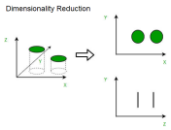
UNSUPERVISED LEARNING TASKS

- Primary tasks within unsupervised learning are:
  1. Clustering
     - grouping similar data points together.
     - it's like sorting objects into categories based on their characteristics.
  2. Association Rules
     - discovers interesting relationships between items in a dataset.
     - it's often used to uncover patterns in transactional data.
  3. Dimensionality Reduction
     - process of reducing the number of features in a dataset while preserving the essential information.
     - it's useful for visualizing high-dimensional data and improving the performance of machine learning models.



Source: https://theappsolutions.com/

Source: https://www.datacamp.com/



Source: https://theappsolutions.com/

- Klasterisasi (*clustering*) digunakan untuk menemukan struktur atau pola dalam kumpulan data yang tidak terkategori.

- Aturan asosiasi (*association rules*) digunakan untuk menemukan hubungan antara berbagai objek dalam satu set, menemukan pola yang sering muncul dalam basis data transaksi, basis data relasional, atau repositori informasi lainnya.

- Reduksi dimensionalitas (*dimensionality reduction*) digunakan untuk mengurangi kompleksitas data sambil mempertahankan bagian-bagian yang relevan dari strukturnya hingga tingkat tertentu.

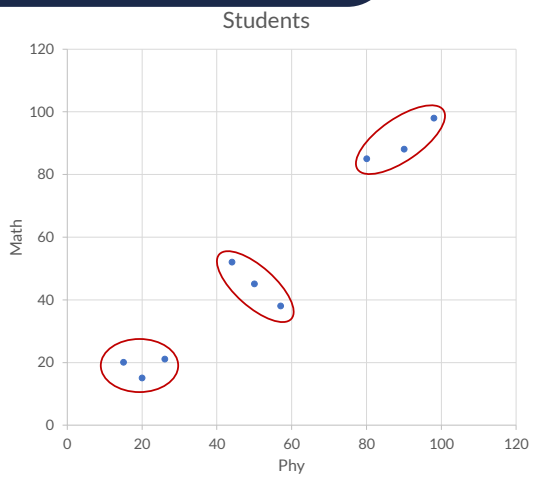
UNSUPERVISED LEARNING TASKS

- Primary tasks within unsupervised learning are:
  1. Clustering applications:
     - Customer Segmentation: Grouping customers based on demographics, purchase history, or behavior.
     - Image Segmentation: Dividing images into regions with similar characteristics.
     - Document Clustering: Grouping similar documents together for information retrieval.
  2. Association Rules applications:
     - Market Basket Analysis: Identifying products that are frequently bought together.
     - Web Usage Mining: Understanding user behavior on websites.
     - Fraud Detection: Identifying unusual patterns in financial transactions.
  3. Dimensionality Reduction applications:
     - Data Visualization: Reducing dimensions to visualize complex data.
     - Feature Engineering: Creating new features that are more informative.
     - Noise Reduction: Removing irrelevant features from the dataset.

CLUSTERING

CASE STUDY

Students	Phy	Math
A	15	20
B	20	15
C	26	21
D	44	52
E	50	45
F	57	38
G	80	85
H	90	88
I	98	98



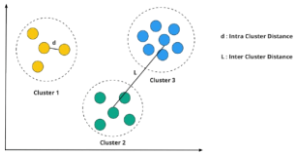
- Anda diminta untuk membagi 9 siswa (A s.d. I) menjadi beberapa kelompok belajar berdasarkan hasil ujian fisika (Phy) dan matematika (Math) mereka.
- Anda kemudian membuat grafik *scatter plot* dengan sumbu X mewakili nilai fisika dan sumbu Y mewakili nilai matematika. Dari grafik tersebut, Anda bisa melihat pola pengelompokkan siswa secara visual.
- Meskipun *scatter plot* memberikan gambaran visual yang intuitif, ada beberapa situasi di mana pendekatan ini tidak efektif:
    1. **Data Berdimensi Tinggi:** Ketika data memiliki lebih dari tiga dimensi, visualisasi menjadi sangat sulit dan bahkan tidak mungkin.
    2. **Dataset Besar:** Untuk dataset yang sangat besar, visualisasi bisa menjadi sangat kompleks dan sulit ditafsirkan.
    3. **Objektivitas:** Bias mungkin terjadi dalam interpretasi visual.

WHAT IS CLUSTERING?

- A cluster is a collection of **data objects** which are
  - **Similar** (or related) **to one another within the same group** (i.e., cluster)
  - **Dissimilar** (or unrelated) **to the objects in other groups** (i.e., clusters)
- Clustering (or cluster analysis, data segmentation, …)
  - Given a set of data points, partition them into a set of groups (i.e., clusters) which are as similar as possible
- Cluster analysis is **unsupervised learning** (i.e., no predefined classes)
  - This contrasts with *classification* (i.e., *supervised learning*)
- Typical ways to use/apply cluster analysis
  - As a stand-alone tool to get insight into data distribution, or
  - As a preprocessing (or intermediate) step for other algorithms

WHAT IS GOOD CLUSTERING?

- A good clustering method will produce high quality clusters which should have
  - **High intra-cluster similarity**: **cohesive** within clusters
  - **Low inter-cluster similarity**: **distinctive** between clusters



- Quality function
  - There is usually a separate "quality" function that measures the "goodness" of a cluster
  - It is hard to define "similar enough" or "good enough"
    - The answer is typically highly subjective
- There exist many similarity measures and/or functions for different applications

APPLICATIONS OF CLUSTER ANALYSIS

- **A key intermediate step for other data mining tasks**
  - Generating a compact summary of data for classification, pattern discovery, hypothesis generation and testing, etc.
  - Outlier detection: Outliers—those "far away" from any cluster
- **Data summarization, compression, and reduction**
  - e.g. Image processing: vector quantization
- **Collaborative filtering, recommendation systems, or customer segmentation**
  - Find like-minded users or similar products
- **Dynamic trend detection**
  - Clustering stream data and detecting trends and patterns
- **Multimedia data analysis, biological data analysis and social network analysis**
  - e.g. Clustering images or video/audio clips, gene/protein sequences, etc.

CONSIDERATIONS FOR CLUSTER ANALYSIS

- **Partitioning criteria**
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable, e.g., grouping topical terms)
- **Separation of clusters**
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- **Similarity measure**
  - Distance-based (e.g., Euclidean, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- **Clustering space**
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

REQUIREMENTS AND CHALLENGES

- **Quality**
  - Ability to deal with different types of attributes: Numerical, categorical, text, multimedia, networks, and mixture of multiple types.
  - Discovery of clusters with arbitrary shape.
  - Ability to deal with noisy data.
- **Scalability**
  - Clustering all the data instead of only on samples.
  - High dimensionality.
  - Incremental or stream clustering and insensitivity to input order.
- **Constraint-based clustering**
  - User-given preferences or constraints; domain knowledge; user queries.
- **Interpretability and usability**
  - The final generated clusters should be semantically meaningful and useful.

CLUSTERING METHODS

- There are many clustering algorithms in the literature.
- In general, the major fundamental clustering methods can be classified into the following categories:
  - **Partitioning** Methods
  - **Hierarchical** Methods
  - **Density-based** Methods
  - **Grid-Based** Methods

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

K-Means

BASIC CONCEPTS

- **Partitioning method**: discovering the groupings in the data by optimizing a specific objective function and iteratively improving the quality of partitions.
  - K-partitioning method: partitioning a dataset $\boldsymbol{D}$ of $\boldsymbol{n}$ objects into a set of $\boldsymbol{K}$ clusters so that an objective function is optimized.
  - A typical objective function: **Sum of Squared Errors (SSE)**, in this case the sum of squared distances is minimized
- Problem definition: Given K, find a partition of K *clusters* that optimizes the chosen partitioning criterion
  - Global optimal: Needs to exhaustively enumerate all partitions
  - Heuristic methods (i.e., greedy algorithms): K-*Means*, K-*Medians*, K-*Medoids*, etc.

SUM OF SQUARED ERRORS (SSE)

- Given a dataset consist of N observations, $X = \{x_1, \dots, x_N\}$
- Its partition Π can be defined as $\Pi = \{C_1, \dots, C_K\}$, where $\forall_{i \neq j} C_i \cap C_j = \emptyset, \cup_{i=1}^{K} C_i = X, \forall_i C_i \neq \emptyset$
- Sum of squared errors (SSE) can be defined as:

$$SSE(X, \Pi) = \sum_{i=1}^{K} \sum_{x_j \in C_i} \|x_j - m_i\|^2$$

  - $\|\cdot\|$ is the Euclidean distance
  - m_i is the centroid of the cluster C_i
- m_i can be computed as the sample mean:

$$m_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j$$

THE K-MEANS CLUSTERING METHOD

- *K-Means* (MacQueen'67, Lloyd'57/'82)
  - Each cluster is represented by the center of the cluster
- Given K, the number of clusters, the *K-Means* clustering algorithm is outlined as follows
    - Select K points as initial centroids
    - **Repeat**
      - Form K clusters by assigning each point to its closest centroid
      - Re-compute the centroids (i.e., *mean point*) of each cluster
    - **Until** convergence criterion is satisfied
- Different kinds of measures can be used
  - Manhattan distance (L_1 norm), Euclidean distance (L_2 norm), Cosine similarity

J. MacQueen, Some Methods for Classification and Analysis of Multivariate Observations, In Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, 1967.
S. Lloyd, Least Squares Quantization, PCM. IEEE Trans. on Information Theory, 28(2), 1982.

EXAMPLE (1/4)

Instances	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

Point	Distance from m_1	Distance from m_2	Cluster Membership
A	2.00	2.24	C1
B	2.83	2.24	C2
C	3.61	2.83	C2
D	4.47	3.61	C2
E	1.00	1.41	C1
F	3.16	2.24	C2
G	0.00	1.00	C1
H	1.00	0.00	C2

1. Tentukan jumlah *cluster*, misalnya $K = 2$
2. Tentukan *centroid* awal secara acak, misalnya $m_1 = (1,1)$ dan $m_2 = (2,1)$
3. Tempatkan tiap objek ke *cluster* terdekat berdasarkan nilai *centroid* yang paling dekat jaraknya, $C_1 = \{A, E, G\}$ & $C_2 = \{B, C, D, F, H\}$

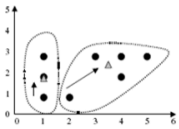
Kalkulasi nilai SSE:

$SSE = (2^2 + 1^2 + 0^2) + (2.24^2 + 2.83^2 + 3.61^2 + 2.24^2 + 0^2) = 36$

EXAMPLE (2/4)

Instances	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

Point	Distance from m_1	Distance from m_2	Cluster Membership
A	1.00	2.67	C1
B	2.24	0.85	C2
C	3.16	0.72	C2
D	4.12	1.52	C2
E	0.00	2.63	C1
F	3.00	0.57	C2
G	1.00	2.95	C1
H	1.41	2.13	C1



4. Berdasarkan *cluster* sebelumnya, hitunglah nilai *centroid* yang baru

$$m_1 = \left(\frac{(1+1+1)}{3}, \frac{(3+2+1)}{3}\right) = (1,2)$$

$$m_2 = \left(\frac{(3+4+5+4+2)}{5}, \frac{(3+3+3+2+1)}{5}\right) = (3.6,2.4)$$

5. Tempatkan kembali setiap objek dengan memakai pusat *cluster* yang baru, $C1 = \{A, E, G, H\}$ & $C2 = \{B, C, D, F\}$

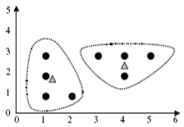
Kalkulasi nilai SSE:

$$SSE = (1^2 + 0^2 + 1^2 + 1.41^2) + (0.85^2 + 0.72^2 + 1.52^2 + 0.57^2) = 7.88$$

EXAMPLE (3/4)

Instances	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

Point	Distance from m_1	Distance from m_2	Cluster Membership
A	1.27	3.01	C1
B	2.15	1.03	C2
C	3.02	0.25	C2
D	3.95	1.03	C2
E	0.35	3.09	C1
F	2.76	0.75	C2
G	0.79	3.47	C1
H	1.06	2.66	C1



4. Berdasarkan *cluster* sebelumnya, hitunglah nilai *centroid* yang baru

$$m_1 = \left(\frac{(1+1+1+2)}{4}, \frac{(3+2+1+1)}{4}\right) = (1.25, 1.75)$$

$$m_2 = \left(\frac{(3+4+5+4)}{4}, \frac{(3+3+3+2)}{4}\right) = (4, 2.75)$$

5. Tempatkan kembali setiap objek dengan memakai pusat *cluster* yang baru, $C1 = \{A, E, G, H\}$ & $C2 = \{B, C, D, F\}$

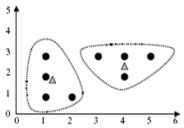
Kalkulasi nilai SSE:

$$SSE = (1.27^2 + 0.35^2 + 0.79^2 + 1.06^2) + (1.03^2 + 0.25^2 + 1.03^2 + 0.75^2) = 6.25$$

EXAMPLE (4/4)

Instances	X	Y
A	1	3
B	3	3
C	4	3
D	5	3
E	1	2
F	4	2
G	1	1
H	2	1

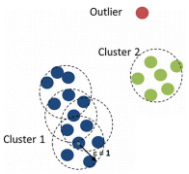
Point	Distance from m_1	Distance from m_2	Cluster Membership
A	1.27	3.01	C1
B	2.15	1.03	C2
C	3.02	0.25	C2
D	3.95	1.03	C2
E	0.35	3.09	C1
F	2.76	0.75	C2
G	0.79	3.47	C1
H	1.06	2.66	C1



- Dapat dilihat pada iterasi 2, tidak ada perubahan anggota lagi pada masing-masing *cluster* (*cluster* yang didapatkan di iterasi 2 sama dengan *cluster* pada iterasi 1).
- Perhatikan pula perbedaan nilai SSE di iterasi 1 (7.88) dan iterasi 2 (6.25) tidak lagi signifikan seperti halnya iterasi 0 (36) dan iterasi 1 (7.88). Maka, proses dihentikan di iterasi 2.
- Hasil akhir yaitu: $cluster_1 = \{A, E, G, H\}$ dan $cluster_2 = \{B, C, D, F\}$ dengan nilai $SSE = 6{,}25$ dan total iterasi 2.

DISCUSSION ON THE K-MEANS METHOD

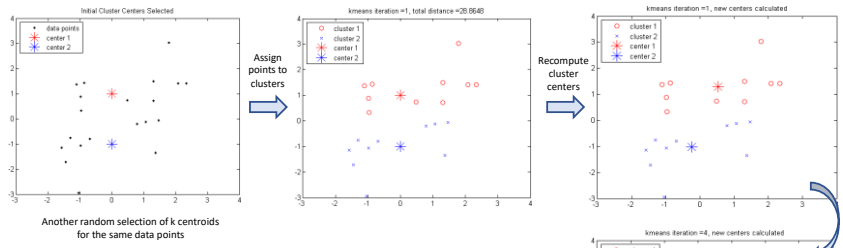
- **Efficiency** (time complexity) : $O(tKn)$ where n: # of objects, K: # of clusters, and t: # of iterations
  - Normally, $K \ll n, t \ll n$; thus, relatively scalable and efficient
- K-means clustering often ***terminates at* a *local optimal***
  - Initialization can be important to find high-quality clusters
- **Need to specify *K***, the *number* of clusters, in advance
  - There are ways to automatically determine the "*best*" *K*
  - In practice, one often runs a range of values and selected the "*best*" *K* value
- **Sensitive to noisy data and *outliers***
  - Variations: using K-medians, K-medoids, etc.
- K-means is applicable only to objects in a continuous n-dimensional space
  - Using the K-modes for ***categorical data***
  - Not suitable to discover clusters with ***non-convex shapes***
  - Using density-based clustering, kernel K-means, etc.



VARIATIONS OF K-MEANS

- There are many variants of the *K-Means* method, varying in different aspects
  - Choosing better initial centroid estimates
    - *K-means++*, *Intelligent K-Means*, *Genetic K-Means*
  - Choosing different representative prototypes for the clusters
    - *K-Medoids*, *K-Medians*, *K-Modes*
  - Applying feature transformation techniques
    - *Weighted K-Means*, *Kernel K-Means*

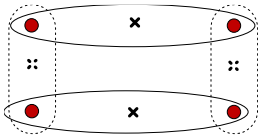
POOR INITIALIZATION LEAD TO POOR CLUSTERING



- Rerun of the *K-Means* using another random *K* seeds
- This run of *K-Means* generates a poor quality clustering

- K-Means seperti permainan mengelompokkan bola-bola berwarna ke dalam beberapa keranjang. Kita ingin mengelompokkan bola-bola sedemikian rupa sehingga bola-bola dalam satu keranjang memiliki warna yang mirip.
- **Kenapa inisialisasi yang buruk bisa menyebabkan hasil klasterisasi yang buruk?**
    - **Kelompok yang Tidak Sempurna:** Bola-bola yang seharusnya berada dalam satu kelompok bisa terpisah ke kelompok lain, atau sebaliknya, bola-bola yang berbeda warna malah tergabung dalam satu kelompok.
    - **Waktu yang Lebih Lama:** Jika awal yang dipilih buruk, algoritma K-Means akan membutuhkan waktu yang lebih lama untuk menemukan kelompok yang benar. Ini karena algoritma harus melakukan banyak percobaan untuk memperbaiki kesalahan awal.

INITIALIZATION PROBLEM AND SOLUTION

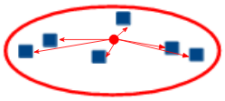


- Different initializations may generate rather different clustering results (some could be far from optimal)
- Original proposal (MacQueen'67): Select K seeds randomly
  - Need to run the algorithm multiple times using different seeds
- There are many methods proposed for better initialization of *k* seeds
  - ***K-Means++*** (Arthur & Vassilvitskii'07):
    - The first centroid is selected at random
    - The next centroid selected is the one that is farthest from the currently selected (selection is based on a weighted probability score)
    - The selection continues until K centroids are obtained

Performance Evaluation of Clustering Algorithms

CLUSTERING EVALUATION METRICS: INERTIA

- Inertia calculates the sum of all the points within a cluster from the centroid of that cluster.
- This distance within the clusters is known as **intra-cluster distance**. So, inertia gives us the sum of intracluster distances:



Intra cluster distance

- The lesser the inertia value, the better the clusters.

CLUSTERING EVALUATION METRICS: DUNN INDEX

- Along with the distance between the centroid and points, **the Dunn index also takes into account the distance between two clusters**.

Intra cluster distance Inter cluster distance

- This distance between the centroids of two different clusters is known as **inter-cluster distance**. Let's look at the formula of the Dunn index:

$$\text{Dunn Index} = \frac{\min(\text{Inter cluster distance})}{\max(\text{Intra cluster distance})}$$

- The more the value of the Dunn index, the better will be the clusters.

SUMMARY

- Unsupervised learning finds hidden patterns or intrinsic structures in data.
- Unsupervised learning mainly uses clustering, association rules, and dimesionality reduction techniques.
- A good clustering method will produce high quality clusters which should haveh high intra-cluster similarity and low inter-cluster similarity.
- K-means algorithm is an iterative algorithm that tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group.
- Inertia and Dunn Index can be used to evaluate a clustering algorithm performance.

REFERENCES

- S. J. Russell and P. Borvig, *Artificial Intelligence: A Modern Approach (4th Edition)*, Prentice Hall International, 2020.
  - Chapter 19. Learning from Examples
- J. Han and M. Kamber, "*Data Mining: Concepts and Techniques (3rd Edition)*," Elsevier, 2012.

eof

Faculty of Informatics and Electrical Engineering