IE403.E31
Khai thác dữ liệu và truyền thông xã hội

Huỳnh Công Bằng

GVHD: ThS. Mai Xuân Hùng

Tp. Hồ Chí Minh, Ngày 30 tháng 7 năm 2025

Mục lục

Course Project Guidelines

For this course, you'll be implementing algorithms using **Python**. A significant part of your assessment will involve **team presentations**, which also require creating a **video report**.

The course spans **eight sessions.** Your **midterm grade** will be based on submitted assignments and attendance. When programming, remember to **integrate a suitable user interface.** All necessary **data will be sourced from the theoretical lessons**, and you should **pay close attention during lectures** to understand data extraction techniques. It's important to **minimize the use of supporting libraries.**

A key challenge in the IT field, as noted, is its **rapid development**. For further reference, you can consult the textbook "Data Mining" by **Prof. Dr. Do Phuc** and slides by **Dr. Nguyen Van Kiet**.

Chương 1

Overview

1.1 What is data mining?

Data mining is the process of discovering valuable patterns and insights from large datasets.

1.2 Benefit of data mining

Data mining provides valuable insights and knowledge that can be used to **support decision-making, predict future trends or outcomes**, and **effectively summarize large datasets**.

1.3 Data mining process

The data mining process typically begins with understanding the **research field** and its objectives, which then guides the **creation of input data**. This raw data then undergoes initial **preprocessing, including cleaning and encoding**, to ensure its quality and suitability for analysis. Following this, **dimensionality reduction** techniques are often applied to simplify the dataset by reducing the number of features or attributes while retaining essential information. The next crucial step involves **selecting the appropriate task for data mining**, such as classification, clustering, or association rule discovery, which in turn informs the **selection of a suitable data mining algorithm**. The chosen algorithm is then applied to the prepared data to **discover hidden knowledge or patterns**. These **found patterns are then evaluated** to assess their validity, novelty, usefulness, and comprehensibility. Finally, the **extracted knowledge is presented** in an understandable format, making it ready for **deployment and practical use** to inform decision-making or achieve specific business objectives.

1.4 From data to decision

- Data: customer data, store data, demographical data, geographical data
- Information: X lives in Z, S is Y years old, X and S moved, W has money in Z

- Knowledge: A quantity Y of product A is used in region Z, Customers of class Y use x % of C during period D
- Decision: Promote product A in region Z, Mail ads to families of profile P, Cross-sell service B to clients C

1.5 Explain

In the realm of data, **data** refers to isolated, raw facts or figures, like "Nguyen Thi Hoa Mai," "student," "Information Technology," or "Database subject." When these individual pieces of data are connected and given context, they transform into **information**. For instance, the statement "Nguyen Thi Hoa Mai is an Information Technology student, and information Technology has a subject called Database" establishes a relationship between these discrete data points. Building on this, **knowledge** represents a deeper understanding derived from the relationships within information. This understanding can exist at two levels. **Level one knowledge** is specific, linking a small group of information, such as: "Nguyen Thi Hoa Mai is an information Technology student, which is why she must learn the Database subject." Moving to a broader perspective, **level two knowledge** identifies patterns or rules across information, like: "If X is an Information Technology student, then X must learn the Database subject." This progression from raw data to actionable knowledge is fundamental in many analytical fields.

1.6 Two fundamental approaches tto data mining

Data mining can generally be approached in two ways. The first is a **verification-oriented approach** where the process begins with **proposing a hypothesis**. The system then focuses on **testing the validity of this hypothesis** through methods like **queries, reports, and statistical analysis.** This approach is driven by existing assumptions or questions that the data is used to confirm or refute.
Conversely, the **discovery-oriented approach** aims to **uncover hidden knowledge or patterns within the database** without prior specific hypotheses. This method is more exploratory, allowing the algorithms to identify novel and potentially unexpected insights directly from the data.

1.7 Data mining techniques

Data mining employs various techniques to extract valuable insights from datasets. One common approach is **Frequent Itemsets and Association Rules**. This technique aims to discover attributes that frequently appear together within a dataset. From these frequent itemsets, **association rules** are generated to identify the likeihood of attributes occurring simultaneously among a set of objects. For example, a rule might state: "If a customer buys X, they will likely also buy Y" (e.g., 66.6% of customers who buy beer also buy squid).
Another powerful technique is **Sequential Pattern Mining**. This mehod focuses on uncovering common sequential patterns that reflect relationships between events in time-oriented databases. It seeks to identify an "X implies Y" relationship, where the occurrence of event X leads to the subsequent occurrence of event Y. An example could be:

"80% of customers who deposit over 80 million VND in savings will seposit an additional 20 million VND three months later." This technique is invaluable for discovering development trends or predicting future behaviors of objects.

Rough Sets (Reduct) are primarily used for **dimensionality reduction** in data classification problems. This helps simplify complex datasets by removing redundant features while preserving essential information.

Data Classification is a supervised learning technique that involves discovering classification rules for a dataset. These rules are used to assign new data points to predefined classes. For instance, patients exhibiting symptoms like coughing, cold, and headache might be classified into the "malaria" disease group.

Finally, **Clustering** is an unsupervised learning process that groups data objects into clusters. The goal is to ensure that objects within the same cluster have a high degree of similarity, while objects in diffent clusters have a low degree of similarity. This technique helps in identifying natural groupings within data without prior knowledge of categories.

Chương 2

Frequent itemset and Association rule

2.1 Association rule

Based on the frequency of data to highlight importance. For example, if 80% of customers who buy beer also buy cigarettes, this can help increase sales.

2.2 Application of Association rules

2.2.1 Understanding customer buying trends

This application focuses on leveraging customer purchasing data to inform business strategies. By analyzing transaction history, we can identify products that are frequently bought together. This knowledge is crucial for optimizing store layouts and managing inventory. For example, if analysis reveals that customers who buy beer often also buy nuts, a store can strategically place these items near each other to boost sales. This also helps in predicting future demand for products. By understanding these purchasing patterns, businesses can make more accurate forecasts for product imports, preventing both overstocking and product shortages.

2.2.2 Analyzing online inventory data

This involves applying association rule mining to e-commerce platforms to enhance the online shopping experience and streamline product management. By examine online sales data, businesses can distinguish between popular, fast-moving items and those with low demand. For instance, if a customer Browse product A frequently also buys product B, the website can be designed to automatically suggest product B, thereby increasing the likeihood of a larger purchase. This data also informs the ongoing management of the online product catalog. It helps in deciding which new products to introduce based on existing buying trends and which underperforming products should be removed or replaced to maintain an engaging and profitable product line.

2.3 How to represent this rule

$Wet\ Wipe \Rightarrow Beer[0.5\%, 60\%]$: This means that in a list of 'n' invoices, we can see that 0.5% of the total invoices contain both beer and wet wipes, and among invoices that include wet wipes, 60% will also contain beer

- Wet Wipe: antecedent
- Beer: consequent
- 0.5%: support
- 60%: confident

2.4 problem statement

In a data mining context, we define the following:

- O is a finite, non-empty set of **transactions** (or invoices)
- I is a finite, non-empty set of **items**
- R is a binary **relation** between O and I. If an element (o, i) belongs to R, where $o \in O$ and $i \in I$, it signifies that transaction o contains item i.

The data mining context is therefore represented as the triplet (O, I, R).

2.5 Example of a data mining context

InvoiceID	ItemID
o1	i1
o1	i2
o1	i3
o2	i2
o2	i3
o2	i4
o3	i2
o3	i3
o3	i4
o4	i1
o4	i2
o4	i3
o5	i3
o5	i4

2.6 Understanding Suport

Given a data mining context (O, I, R) and an itemset S ($S \subset I$), the **support** of S is defined as the ratio of the number of transactions that contain S to the total number of transactions in O.
It's formally denoted as:

$$SP(S) = \frac{|\rho(S)|}{|O|}$$

Where $\rho(S)$ represents the set of transactions that contain all items in S.

2.7 Frequent itemset

They are itemsets whose **support** is greater than or equal to a predefined threshold, known as *minsupp*.

2.8 How to find frequent itemsets: An Example Walk-through

2.8.1 Question

Given a data mining context, find all **frequent itemsets** that satisfy a minimum support threshold (*minsupp*) of **0.4**.

InvoiceID	ItemID
o1	i1
o1	i2
o1	i3
o2	i2
o2	i3
o2	i4
o3	i2
o3	i3
o3	i4
o4	i1
o4	i2
o4	i3
o5	i3
o5	i4

2.8.2 Answer

Since the minimum support(*minsupp*) is 0.4 and the total number of transactions is 5, any **frequent itemset** satisfying the problem's requirements must appear in at least $0.4 \times 5 = 2$ transactions

Establish a binary matrix

	$i1$	$i2$	$i3$	$i4$
$o1$	1	1	1	0
$o2$	0	1	1	1
$o3$	0	1	1	1
$o4$	1	1	1	0
$o5$	0	0	1	1

Identifying frequent itemsets that meet the minimum support threshold

The candidate itemsets of size 1 are $F_1 = \{\{i1\}, \{i2\}, \{i3\}, \{i4\}\}$

- $SP(\{i1\}) = \frac{2}{5} = 0.40$; Popular ($\geq minsupp$)
- $SP(\{i2\}) = \frac{4}{5} = 0.80$; Popular ($\geq minsupp$)
- $SP(\{i3\}) = \frac{5}{5} = 1.00$; Popular ($\geq minsupp$)
- $SP(\{i4\}) = \frac{3}{5} = 0.60$; Popular ($\geq minsupp$)

The frequent itemsets consisting of a single item are: $C1 = \{\{i1\}, \{i2\}, \{i3\}, \{i4\}\}$

Frequent itemsets using Apriori's Pruning Strategy

The **candidate set** C_k is generated by joining L_{k-1} with itself. This is the **join step**. The **pruning step** (or "trimming step") then applies the Apriori principle: any itemset of size $(k-1)$ that is not frequent cannot be a subset of a frequent itemset of size k

Identifying frequent itemsets that meet the minimum support threshold (to be continued)

The candidate itemsets of size 2 from the frequent itemsets $C1$ are

$$F_2 = \{\{i1, i2\}, \{i1, i3\}, \{i1, i4\}, \{i2, i3\}, \{i2, i4\}, \{i3, i4\}\}$$

- $SP(\{i1, i2\}) = \frac{2}{5} = 0.40$; Popular ($\geq minsupp$)
- $SP(\{i1, i3\}) = \frac{2}{5} = 0.40$; Popular ($\geq minsupp$)
- $SP(\{i1, i4\}) = \frac{0}{5} = 0.00$; Not popular ($< minsupp$)
- $SP(\{i2, i3\}) = \frac{4}{5} = 0.80$; Popular ($\geq minsupp$)
- $SP(\{i2, i4\}) = \frac{2}{5} = 0.40$; Popular ($\geq minsupp$)
- $SP(\{i3, i4\}) = \frac{3}{5} = 0.60$; Popular ($\geq minsupp$)

The frequent itemsets consisting of two items are:

$$C2 = \{\{i1, i2\}, \{i1, i3\}, \{i2, i3\}, \{i2, i4\}, \{i3, i4\}\}$$

The candidate itemsets of size 3 from the frequent itemsets $C2$ are

$$F_3 = \{\{i1, i2, i3\}, \{i1, i2, i4\}, \{i1, i3, i4\}, \{i2, i3, i4\}\}$$

- $SP(\{i1, i2, i3\}) = \frac{2}{5} = 0.40$; Popular ($\geq minsupp$)
- $SP(\{i2, i3, i4\}) = \frac{2}{5} = 0.40$; Popular ($\geq minsupp$)
- $\{i1, i4\}$is not popular, therefore $\{i1, i2, i4\}$ and $\{i1, i3, i4\}$ are not popular

The frequent itemsets consisting of three items are:

$$C3 = \{\{i1, i2, i3\}, \{i2, i3, i4\}\}$$

The candidate itemsets of size 4 from the frequent itemsets $C3$ are $F_4 = \{\{i1, i2, i3, i4\}\}$

- $\{i1, i4\}$is not popular, therefore $\{i1, i2, i3, i4\}$is not popular

Frequent itemsets that meet the minimum support threshold are $\{i1\}$, $\{i2\}$, $\{i3\}$, $\{i4\}$, $\{i1, i2\}$, $\{i1, i3\}$, $\{i2, i3\}$, $\{i2, i4\}$, $\{i3, i4\}$, $\{i1, i2, i3\}$, $\{i2, i3, i4\}$