Summary of Theory 2: Introduction to Data Mining

Based on Lecture Notes by Fernando Lucas Bação, NOVA IMS (2025/2026)

1 Growth of the Digital Universe

The digital universe is expanding exponentially, leading to a data deluge. Organizations generate and store vast amounts of information daily, creating both opportunities and challenges for extracting valuable insights.

2 Big Data

Definition

Big Data refers to datasets that are too large or complex for traditional data-processing systems, requiring new technologies and analytical methods.

Characteristics (3Vs)

- **Volume:** Enormous quantities of data generated continuously.
- **Velocity:** High speed of data generation and processing in real-time.
- **Variety:** Different data types — structured, semi-structured, and unstructured.

3 Artificial Intelligence (AI)

AI is the science of creating systems that exhibit human-like intelligence. It enables machines to perform tasks that typically require human cognition, such as reasoning, perception, and learning. The scope of AI evolves over time as technology advances.

4 Machine Learning (ML)

Machine Learning is a subfield of AI that focuses on algorithms capable of learning from data without explicit programming. ML systems detect patterns and improve performance through experience.

- Learns automatically from data.
- Builds predictive or descriptive models based on examples.
- Foundation for modern AI applications.

5 Data Science

Data Science studies how to extract knowledge and insight from data. It integrates statistics, computer science, and domain expertise.

Roles in Data Science

- **Business Analyst:** Bridges business understanding with data insights.
- **Data Scientist:** Designs and implements models to extract knowledge.
- **Data Engineer:** Manages data infrastructure and pipelines ensuring data quality and accessibility.

6 Building Models

Model development involves selecting appropriate algorithms and defining the relevant features that describe the problem space. Feature engineering is crucial for model performance.

What is a Feature and an attribute?

An attribute is a data field that represents a characteristic or feature of a data object.
The nouns attribute, dimension, feature, and variable are often used interchangeably in the literature.
The term dimension is commonly used in data warehouses.
The machine learning literature tends to use the term feature, whereas statisticians prefer the term variable.
Data mining and database professionals commonly use the term attribute.
Attributes describing a customer object can include, for example, the customer ID, name, and address. Observed values for a given attribute are known as observations. A set of attributes used to describe a given object is called an attribute vector (or feature vector).

Feature Construction

- **ETL Process:** Extract, Transform, and Load data.
- Create new attributes such as recency, frequency, monetary value, or behavioral indices.
- Features must capture the key properties of the phenomena being studied.

7 Relevance of Data

The choice of data and features directly affects model accuracy. Domain knowledge is vital to select meaningful variables and to avoid irrelevant or misleading features.

8 Statistics and Data Science

Statistics provides the theoretical foundation for inference and hypothesis testing, while Data Science extends this with computational and algorithmic approaches to large and complex datasets.

9 Canonical Tasks in Data Mining

- **Predictive Modelling (Supervised Learning):** Builds models to predict outcomes for new data.
- **Descriptive Modelling (Unsupervised Learning):** Summarizes data to discover hidden patterns.

10 Supervised Learning

Uses labeled data to train models for classification or regression. Examples include fraud detection and image recognition.

11 Unsupervised Learning

Finds structures or relationships in unlabeled data, such as clustering, association rules, and data visualization.

12 Class / Concept Description: Characterization and Discrimination

General idea

Data entries can be associated with *classes* or *concepts* (for example: product categories like *computers* or customer concepts like *bigSpenders*). A concise and precise description of this class or concept is called a **class/concept description**. These descriptions can be derived by:

1. **Data characterization:** summarizing the data of the *target class* in general terms;
2. **Data discrimination:** comparing the target class with one or more *contrasting classes*; or
3. **Both:** combining characterization and discrimination.

Data characterization

Data characterization is a summarization of the general characteristics or features of a target class. Typically, data for the target class are collected through a query. Examples of methods and presentation forms:

- **Methods:** statistical summaries (means, counts), plots, attribute-oriented induction, OLAP roll-up.
- **Output forms:** tables, bar charts, pie charts, curves, multidimensional cubes, crosstabs, generalized relations, or *characteristic rules.*

Illustrative example: Summarize customers who spend more than $5000 per year. Result: a profile that may state they are 40–50 years old, employed, and have excellent credit ratings; the user should be able to drill down by dimensions such as occupation.

Data discrimination

Data discrimination compares general features of the target class against one or more contrasting classes. It uses similar methods as characterization but emphasises comparative measures to highlight distinguishing properties.

- **Use case:** Compare software products with sales up by at least 10% versus products with sales down by at least 30%.
- **Output:** discriminant rules, comparative charts, discriminative measures.

13 Explain like I'm 5

Toy-store analogy: Imagine you have a big toy store:

- A **class/concept** is a named box (e.g. "bigSpenders").
  For toys: classes could be "cars" and "dolls."
  For kids: classes could be "big spenders" (buy lots of toys) and "budget spenders" (buy only one).
- **Characterization** = describing what is inside the box For example: You look at the "big spenders" box and say,

  "These kids are usually 8 years old, love robots, and come to the store every week."

  You're not comparing them to anyone else — you're just describing them. That's called characterization — you tell what something is like.
- **Discrimination** = comparing two boxes to see differences ("big spenders buy robots, budget spenders buy coloring books").

14 Are classes only for supervised learning?

- **Supervised learning:** classes are labels used for training (labeled data). Example: *spam* vs *not spam.*
- **Characterization/discrimination:** classes can be *user-defined groups* (e.g., customers who spent ¿ $5000). These groups are used to describe or compare data and are not necessarily used to train a predictive model.

15 Difference between "classes" and "categories"

- **Category:** the variable or concept/feature (e.g., "Fruit type").
- **Class:** a specific value within that category (e.g., "Apple", "Orange").
- **Instance:** a single data record that belongs to one class.

Mnemonic: *Category = the question; Class = the answer.*

16 The Data Mining Process

The process follows structured methodologies like KDD (Knowledge Discovery in Databases) or CRISP-DM (Cross-Industry Standard Process for Data Mining), involving:

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment