Introduction to Data Mining and Machine Learning Techniques

Iza Moise, Evangelos Pournaras
Overview

Main principles of data mining
  Definition
  Steps of a data mining process
  Supervised vs. unsupervised data mining

Applications

Data mining functionalities
Definition

**Data mining**

is the **automated** process of discovering **interesting** (non-trivial, previously unknown, insightful and potentially useful) information or **patterns**, as well as **descriptive**, understandable, and **predictive** models from (large-scale) data.

Also known as:

- Knowledge discovery in databases (KDD)
- Data analysis
- Information harvesting
- Business intelligence
Goals

✓ search consistent patterns and/or systemic relationships between data

✓ validate the findings by applying the detected patterns to new subsets of data

✓ predict new findings on new datasets
Data Mining is...

- k-means clustering
- decision trees
- neural networks
- Bayesian networks
Data Mining is not...

- Data warehousing
- SQL / Ad Hoc Queries / Reporting
- Software Agents
- Online Analytical Processing (OLAP)
- Data Visualization
Knowledge Discovery Process
Steps of a data mining process

1. Exploration
2. Model building and validation
3. Deployment → prediction
Exploration

- data preparation
  - data cleaning, data transformation etc.
- elaborate exploratory analyses using a wide variety of graphical and statistical methods, depending on the nature of the analytic problem
- determine the general nature of models that can be taken into account in the next stage
Model building and validation, and Deployment

- Model building and validation:
  - consider various models and choose the best one
  - validate the model on existing data

- Deployment:
  - apply the model to new data
Model building and validation, and Deployment

- Model building and validation:
  - consider various models and choose the best one
  - validate the model on existing data

- Deployment:
  - apply the model to new data
Model building and validation, and Deployment

- **Model building and validation:**
  - consider various models and choose the best one
  - validate the model on existing data

- **Deployment:**
  - apply the model to new data
Two styles of data mining: Predictive and Descriptive

1. **Predictive**
   → Predict the value of a specific attribute (target or dependent) based on the value of other attributes (explanatory or predictors)

2. **Descriptive**
   → Derive patterns that summarise the relationships among data points
Supervised vs. Unsupervised

**Supervised**

- *predictive or directed*
- useful when you have a specific target value to predict about your data
- Classification, Regression, Anomaly Detection

**Unsupervised**

- *descriptive or undirected*
- finds hidden structure and relation within the data
- Clustering, Association, Feature Extraction
Supervised Data Mining

- a pre-specified target variable
- explanatory variables or predictors and ↔ one (or more) dependent variables or target

- goal: specify relationships between predictors and target
- training: the model is given many training data where the target is known
- testing: the model is applied to data for which the target is unknown
Unsupervised Data Mining

- determine the existence of classes or clusters in the data
- *exploratory* analysis
- all variable are treated in the same way
Overview

Main principles of data mining
  Definition
Steps of a data mining process
Supervised vs. unsupervised data mining

Applications

Data mining functionalities
Fielded Applications

- Web mining
  → PageRank, Search Engine, Recommenders, Social Networks
- Screening Images
- Diagnosis
- Marketing and Sales
  → Market basket analysis, Direct marketing
- Biology, biomedicine, astronomy, chemistry
Beers and Diapers
Overview

Main principles of data mining
  Definition
  Steps of a data mining process
  Supervised vs. unsupervised data mining

Applications

Data mining functionalities
1. Supervised Data Mining
   ✓ Classification
   ✓ Regression
   ✓ Outlier detection
   ✓ Frequent pattern mining

2. Unsupervised Data Mining
   ✓ Clustering
   ✓ Feature Extraction

→ definition
→ real use-cases
→ method
→ pros and cons
1. Supervised Data Mining
   ✓ Classification
   ✓ Regression
   ✓ Outlier detection
   ✓ Frequent pattern mining

2. Unsupervised Data Mining
   ✓ Clustering
   ✓ Feature Extraction

→ definition
→ real use-cases
→ method
→ pros and cons