Data Mining

Guide to AI and Data Courses · Part IV: Data

Knowledge discovery | Association rules | Clustering | Outliers | Interestingness measures | Neighboring courses

Sercan KÜLCÜ · Giresun University · Computer Engineering

© Sercan Külcü. All rights reserved. · Interactive HTML version of these slides: https://sercankulcu.github.io/files/ai_courses/slides/Chapter_25_Data_Mining.html

INTRODUCTION

What Question Does It Ask?

> **How are hidden, useful patterns found in large piles of data?**

Data mining studies how to **discover patterns** in large data collections that are previously unknown, valid and useful. It uses the same algorithms as machine learning, but the goal is not prediction on new examples; it is finding rules, clusters and anomalies in the data itself.

At a glance

APPROACH discovery-driven analysis	**INPUT** large tables or transaction data	**OUTPUT** rules, clusters, outliers
SUCCESS MEASURE interestingness, support, confidence	**MATH LOAD** medium	**DATA NEEDS** large

A Short History

1967
MacQueen publishes the clustering method he names "k-means".

1989
The first workshop on knowledge discovery in databases (KDD) is held; the field gets its name.

1994
Agrawal and Srikant popularize association rule mining with the Apriori algorithm.

1996
Fayyad and colleagues define the KDD process; density-based DBSCAN is published the same year.

2000
CRISP-DM 1.0 is released: an industry-neutral process model for data mining projects.

2006
The Netflix Prize opens: one million dollars for improving recommendation accuracy by 10 percent.

THE COURSE

Core Concepts

Knowledge discovery (KDD)
Selection, preprocessing, transformation, mining, interpretation.

Frequent itemsets
Products often bought together: Apriori, FP-growth.

Association rule
If X then Y; measured by support, confidence and lift.

Clustering
k-means, hierarchical, DBSCAN; unlabeled groups.

Outlier
A record unlike the rest: an error or fraud?

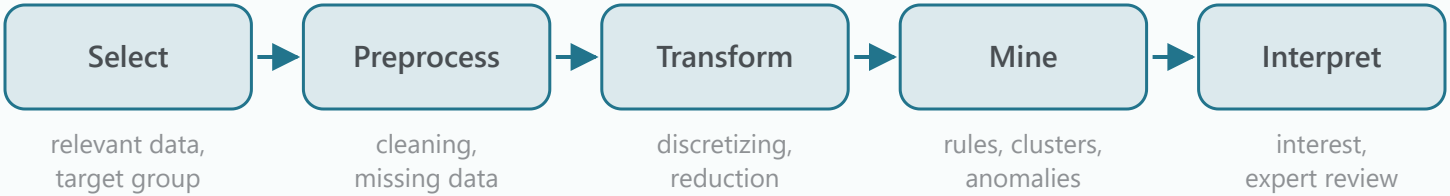
Descriptive model
Readable, explanatory models such as decision trees.

Preprocessing
Missing values, noise, discretization, normalization.

Sequential pattern
Event sequences that recur in time order.

THE COURSE

A Typical Workflow



Example: A supermarket selects a year of receipts, removes returns and bad records, turns each receipt into an itemset, finds rules such as "customers who buy pasta also buy sauce" with Apriori, and presents the high-lift rules to the manager for shelf layout.

THE COURSE

What Does the Course Cover?

1. The data mining process: KDD and CRISP-DM
2. Getting to know data: attribute types, similarity
3. Preprocessing: cleaning, integration, reduction
4. Data warehouses and OLAP cubes
5. Frequent itemsets: Apriori and FP-growth
6. Association rules and interestingness measures
7. Classification: decision trees, rule-based methods
8. Evaluating classifiers and imbalanced classes
9. Clustering: k-means and hierarchical methods
10. Density-based clustering and cluster validity
11. Outlier detection
12. Sequential pattern, text and graph mining

Skills you will gain

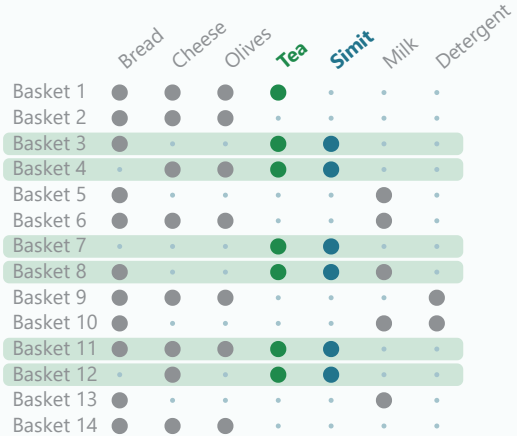
- data preparation with pandas and SQL
- mining association rules with mlxtend
- quick experiments in Weka or Orange
- measuring and interpreting clusters
- translating findings for the business

A typical semester; the order and coverage vary between universities.

TRY IT

Hidden Rules in the Shopping Basket

A corner shop looks at 14 receipts: nobody asked a question; let the data tell which items are bought together. Change the thresholds and Apriori finds the rules again; click a rule to see it in the baskets.



Min support: **3 (21%)** baskets

Min confidence: **60%**

Apriori: size 1: 7 candidates → **6** frequent; size 2: 15 candidates → **10** frequent; size 3: 7 candidates → **4** frequent. **18** rules pass the confidence threshold; 4 of them:

{Simit} → {Tea}
support 43% · confidence 100% · lift 2.00

{Bread, Cheese} → {Olives}
support 43% · confidence 100% · lift **2.00**

{Bread, Simit} → {Tea}
support 21% · confidence 100% · lift **2.00**

{Olives} → {Bread}
support 43% · confidence 86% · lift **1.09**

Data mining is **discovery**: frequent patterns are sought without asking a question first. A rule with lift near 1 is empty (Bread is in everyone's basket anyway); a useful rule has lift clearly above 1.

CONTEXT

Prerequisites

Courses worth knowing before this one, and why they are needed.

Course	Why it is needed
Probability and Statistics	support, confidence, significance
Databases and SQL	tables, joins and grouping: where the data comes from
Data Structures	hash tables and trees: building blocks of algorithms such as FP-growth
Algorithms	complexity: understanding the exponential growth of candidate sets

CONTEXT

Are You Ready for This Course?

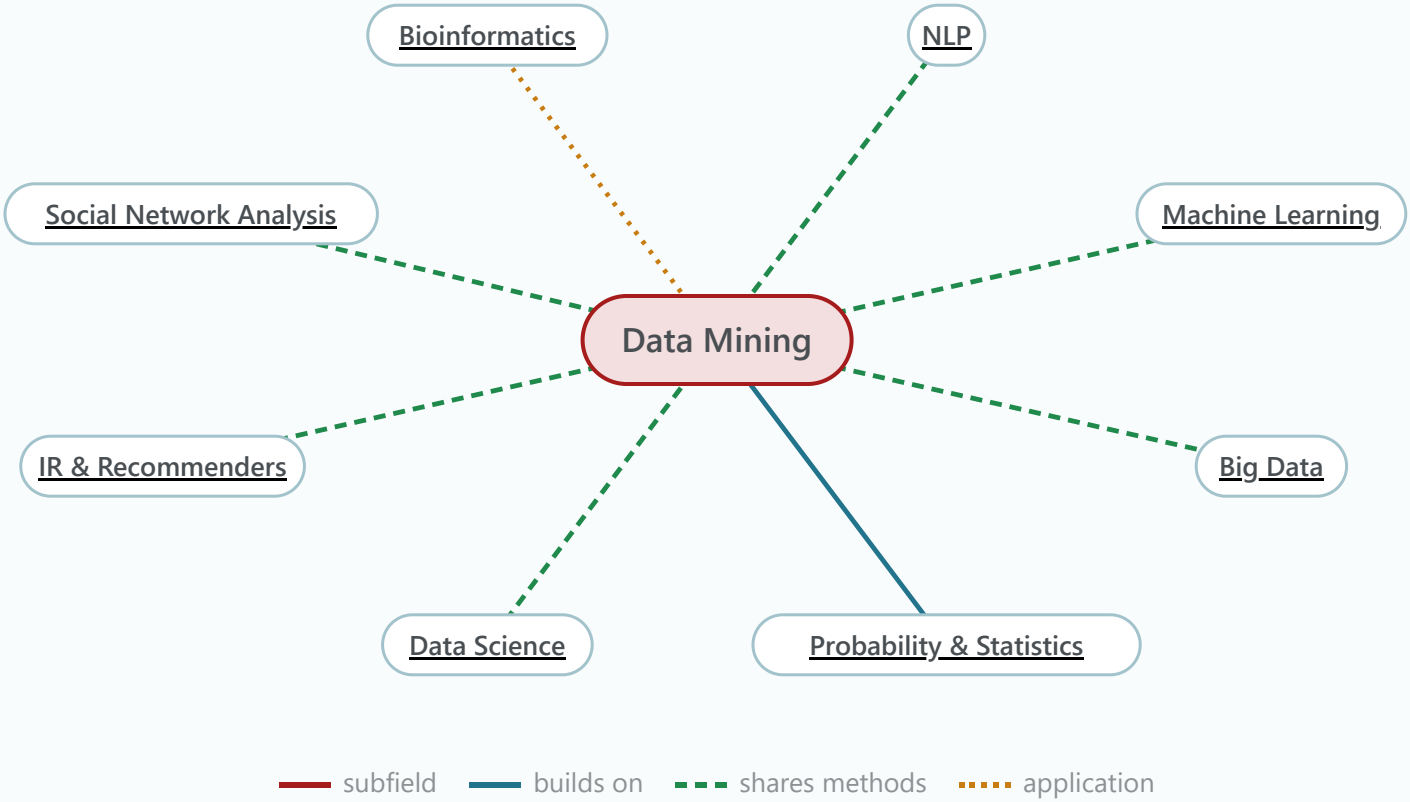
Ask yourself; for every "no", the right-hand column shows where to start.

Ask yourself	If not, start with
Can you join and group two tables in one SQL query?	SQL from a database course: JOIN and GROUP BY
Can you explain conditional probability and independence with an example?	Probability and Statistics
Can you explain when a hash table beats a tree?	A data structures course: hash tables and trees
Can you read a CSV file in Python and count its missing values?	An intro programming course, then a short pandas tutorial

If you can say "yes" to most of these, you are ready; one or two gaps can be closed during the semester.

CONTEXT

Neighbouring Courses



Overlapping Courses

Relation	Course	Shared ground
builds on	This course builds on Probability and Statistics	support, confidence, significance
application	Bioinformatics applies the methods of this course	sequence motifs, clustering
shares methods	Shares methods with Machine Learning	same algorithms; the goal is discovery, not prediction
shares methods	Shares methods with Natural Language Processing and Text Mining	text classification, topic modeling
shares methods	Shares methods with Data Science	exploratory data analysis
shares methods	Shares methods with Big Data Systems	scalable mining
shares methods	Shares methods with Information Retrieval and Recommender Systems	association, collaborative filtering
shares methods	Shares methods with Social Network Analysis	community detection

DIFFERENCES

Where It Differs (1/2)

Criterion	Data Mining	Machine Learning	Data Science
Core question	How are hidden, useful patterns found in large piles of data?	How does a program generalize from examples?	How do we ask the right questions of data and turn the answers into decisions?
Approach	discovery-driven analysis	statistical learning	question, data, analysis, story
Input	large tables or transaction data	labeled or unlabeled examples	raw, messy data
Output	rules, clusters, outliers	a model that makes predictions	report, visualization, model
Success measure	interestingness, support, confidence	error on unseen data	impact on decisions, accuracy
Math load	medium	high: linear algebra, probability, optimization	medium: statistics
Data needs	large	moderate to large	moderate to large

- **Machine Learning:** They share algorithms; ML aims at prediction on new examples, data mining at discovery in the data at hand.
- **Data Science:** Data science covers the whole path from question to decision; data mining focuses on the pattern-finding step.

Where It Differs (2/2)

Criterion	Data Mining	Big Data	Probability & Statistics
Core question	How are hidden, useful patterns found in large piles of data?	How is data that does not fit on one machine stored and processed?	How is uncertainty measured, and how do we draw reliable conclusions from limited data?
Approach	discovery-driven analysis	distributed storage and computation	probability models and inference
Input	large tables or transaction data	huge, fast-flowing data	sample data
Output	rules, clusters, outliers	processed data, query results	estimate, confidence interval, test result
Success measure	interestingness, support, confidence	scalability, latency, cost	unbiasedness, power, significance
Math load	medium	low	the theory itself
Data needs	large	very large	a sample

- **Big Data:** Big data systems store and process data in a distributed way; data mining asks which patterns to look for in it.
- **Probability & Statistics:** Statistics tests a hypothesis set in advance; data mining generates hypotheses from data, which must then be tested.

PRACTICE

Applications

Market basket
Items bought together, shelves and promotions.

Customer segmentation
Grouping similar customers.

Fraud detection
Flagging unusual transactions.

Web usage
Visitors' navigation patterns.

Failure logs
Error codes that occur together.

Medical records
Diagnoses and drugs that occur together.

PRACTICE

Risks and Ethics

Inferring sensitive facts

Pregnancy, illness or money trouble can be inferred from a shopping basket, even if the person never shared it.

Anonymous data is not anonymous

Netflix Prize data was linked to IMDb ratings to identify some users; the planned second contest was cancelled over it.

Data dredging

Search for enough patterns and chance ones look significant; multiple comparison corrections and testing on fresh data are needed.

Discriminatory profiling

Customer segments built from postcodes or shopping patterns can indirectly give some groups different prices and service.

In depth: Explainable AI and AI Ethics.

PRACTICE

Tools and Datasets

Tools and libraries

Weka | KNIME | Orange | RapidMiner | scikit-learn | mlxtend | R (arules)

Well-known datasets

Groceries (arules) | Online Retail (UCI) | Instacart Market Basket | Adult (income) | KDD Cup 1999 | Mushroom

PRACTICE

Common Misconceptions

Common belief	In fact
Data mining is collecting data.	The name misleads: what is mined is knowledge, not data; the data already exists.
Every rule found is useful.	Thousands of rules come out; most are obvious or accidental. Interestingness measures and expert review are needed.
A strong association means causation.	Occurring together does not mean one causes the other; there may be a common cause.
Clustering has one right answer.	The result depends on the number of clusters, distance and scaling; different choices, different clusters.

NEXT STEPS

After This Course

Courses that build on or extend this one

Bioinformatics

Related careers

Data mining specialist | Data analyst | Business intelligence specialist | Customer analytics specialist | Fraud analyst

Data mining is the discovery side of the data courses. From here you can move to recommender systems through rules, to social network analysis through community detection, to big data systems through scalable mining, or to machine learning to focus on prediction.

NEXT STEPS

Further Reading

- Han, Kamber and Pei, *Data Mining: Concepts and Techniques*, 3rd ed., Morgan Kaufmann, 2011.
- Tan, Steinbach, Karpatne and Kumar, *Introduction to Data Mining*, 2nd ed., Pearson, 2019.
- Witten, Frank, Hall and Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed., Morgan Kaufmann, 2017.
- Leskovec, Rajaraman and Ullman, *Mining of Massive Datasets*, 3rd ed., Cambridge University Press, 2020.

The End

Chapter 25: Data Mining

← Previous chapter · Map of the field · Next chapter →