Introduction IE500 Data Mining





- Dr. Sven Hertling
 - Substitute Professor for Data Science
- Research Interests:
 - Knowledge Graph Integration
 - KGs in combination with Large Language Models
 - Information Extraction
- Room: B6 26, B0.21
- eMail: sven.hertling@uni-mannheim.de
- Will teach the lectures





- M.Sc. Andreea Iana
 - Graduate Research Associate
- Research Interests:
 - Recommender Systems
 - Information Retrieval
 - Natural Language Processing
- Room: B6 26, B 0.18
- eMail: andreea.iana@uni-mannheim.de
- Will teach one of the exercise groups and will supervise student projects



- M.Sc. Franz Krause
 - Graduate Research Associate
- Research Interests:
 - Machine Learning Applications on Linked Data
 - Dynamization of Knowledge Graph Embeddings
 - Knowledge Graph Application and Implementation in Industrial Settings
 - Applied Graph Theory
- Room: B6 26, B 0.02
- eMail: <u>franz.krause@uni-mannheim.de</u>
- Will teach one of the exercise groups and will supervise student projects





- M. Sc. Aaron Steiner
 - Graduate Research Associate
- **Research Interests:**
 - Entity Matching using LLMs
 - Fine-tuning LLMs
 - Data Integration
- Room: B6 26, C 1.04
- eMail: aaron.steiner@uni-mannheim.de
- Will teach one of the exercise groups and will supervise student projects





Course Organization



- Registration
 - you have registered via Portal2
 - and been added to ILIAS



- Lecture
 - Time: Wednesday, 10:15 11:45
 - Location: Room B144 Building A 5,6 Part B
 - introduces the principle methods of data mining
 - discusses how to evaluate the learned models
 - presents practical examples of data mining applications

Course Organization - Material



Course Webpage

- <u>https://www.uni-mannheim.de/dws/teaching/course-details/courses-for-master-candidates/ie-500-data-mining</u>
- Provides up-to-date information, lecture slides, video lectures
- ILIAS eLearning System
 - <u>https://ilias.uni-mannheim.de/</u>
 - Exercises
 - Mailing lists, discussion forum,
 - Team project (submission, coaching sessions)



Data Mining (FSS 2019)

Goals and Principles of Data Minir

and providing input for deriving support systems. The course will cover the following tonic



irse provides an introduction to advanced data analysis techniques as a basis for analyzing business dat.

Course Organization - Material



- Online Lecture
 - Part of the course
 - Exercise / Exam relevant



Week	Wednesday (Offline Lecture)	Online Lecture	Thursday (Exercise)
10.02.2025	no lecture		Introduction to Python (13:45-15:15)
17.02.2025	Introduction to Data Mining		Intro
24.02.2025	Preprocessing		Preprocessing
03.03.2025	Classification 1	Nearest Centroids	Classification 1
10.03.2025	Classification 2	Comparing Classifiers	Classification 2
17.03.2025	Regression	Ensembles	Regression
24.03.2025	Clustering and Anomalies	Hierarchical Clustering	Clustering
31.03.2025	Feedback on project outlines	Time Series	Time Series
07.04.2025	Association Analysis and Subgroup Discovery		Association Analysis
14.04.2025	Easter Break		Easter Break
21.04.2025	Easter Break		Easter Break
28.04.2025	Project feedback session	Multi Modal Data	Public holiday
05.05.2025	Project feedback session		Project Work
12.05.2025	Project feedback session		Project Work
19.05.2025	Project Work		Project Presentations
26.05.2025	Q&A		Public holiday

Course Organization - Exercise



- Exercise Groups
 - Students experiment with data sets using Python
- Time and Location (same content only attend **one**):
 - Thursday, 10.15 11.45, A104 Building B6, 26 Part A (Aaron)
 - Thursday, 12.00 13.30, A104 Building B6, 26 Part A (Andreea/Franz)
 - Thursday, 13.45 15.15, A104 Building B6, 26 Part A (Andreea/Franz)



Introduction to Python



- Already last week Thursday 13:45-15:15
- Topics:
 - Setup of environment (Anaconda, Jupyter Notebooks)
 - Python Intro / Design Goals
 - Basic programming concepts in Python
- Support
 - Help with environment setup
 - Q&A
- Material
 - Tutorial slides available on website





Usage of ChatGPT



- We will be using ChatGPT in the exercise to
 - Discuss suitable methods and parameter settings for different use cases
 - Generate and debug Python code for experimenting with the methods



Course Organization - Project



- Project Work
 - Teams of **six** students realize a data mining project
 - Teams may choose their own data sets and tasks
 (in addition, we will propose some suitable data sets and tasks)
 - Write summary about project and present the results
- Deadlines
 - Team formation Sunday, March, 16th, 23:59
 - Submission of project proposal
 - Sunday, March, 23rd, 23:59
 - Submission of final project work report
 - Sunday, May 18th, 23:59
 - Submission of Presentation (PDF)
 - Wednesday, May 21st, 23:59



Course Organization - Exam



- Date and Time: Thursday, 12th June 2025
- Duration: 60 minutes
- Structure: 6 open questions that
 - Check whether you have understood the lecture content
 - We try to cover all major chapters of the lecture
 - Require you to describe the ideas behind algorithms and methods
 - Often: How do methods react to special patterns in the data?
 - Might require you to do some simple calculations for which
 - You need to know the most relevant formulas
 - You do not need a calculator

Course Organization - Exam



- There is only one exam per semester
 - Because course is offered every semester
 - The next exam date is at the end of the upcoming FSS
 - i.e., no retake date!
- Upon failure, you will have to redo both the project and the exam in another semester
 - Unfortunately, we cannot carry over your project mark
- Final grade
 - 75 % written exam, 20% project report, 5% project presentation

Schedule

 All updates will be made available at the course homepage





Textbooks for the Course



 Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining. 2nd Edition.
 Pearson / Addison Wesley.

• Aurélien Géron:

Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow. 2nd or 3rd Edition, O'Reilly, 2019 or 2022

• Scikit-learn Documentation:

https://scikit-learn.org/stable/user_guide.html



Videos and Screencasts



• Videos

- <u>https://www.uni-mannheim.de/dws/teaching/lecture-videos</u> (VPN!)
- Lecture Videos By Heiko Paulheim (HWS 2020) and Christian Bizer (FSS 2020)
- Screencasts for the Exercises by Ralph Peeters (FSS 2022)
- Keep in mind, that the lecture and exercise change over time



Questions?





What is Data Mining?



- Large quantities of data are collected about all aspects of our lives
- This data contains interesting patterns
- Data Mining helps us to
 - 1. Discover these patterns and
 - 2. Use them for decision making across all areas of society, including
 - Business and industry
 - Science and engineering
 - Medicine and biotech
 - Government
 - Individuals





Routers Disks Tape Drives Cores 12.1 к 357.9 к 220.0 к 180 304 **5.1** к

https://home.cern/news/news/computing/new-data-centre-cern http://cern.ch/go/datacentrebynumbers

• CERN

- Large Hadron Collider
 - 45 petabytes per week produced (February 2024)
- 820 petabytes of data archived on tape
- 1005 petabytes of disk space available (August 2024)
- Discover
 - Patterns in the experiments





Facebook

- 4 Petabyte of new data generated every day
- over 300 Petabyte in
 Facebook's data
 warehouse
- Predict
 - Interests and behavior of over one billion people

https://www.brandwatch.com/blog/facebook-statistics/ http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/





- US Library of Congress
 - $\approx 235 \text{ TB}$ archived

- Discover
 - Topic distributions*
 - Citation networks
- Train
 - Large Language Models

https://www.brandwatch.com/blog/facebook-statistics/ http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/





University of Mannheim | IE500 Data Mining | Introduction and Course Outline | Version 1.09.2024 https://www.domo.com/learn/infographic/data-never-sleeps-12

"We are Drowning in Data... but starving for knowledge!"



- We are interested in the patterns, not the data itself!
- Data Mining methods help us to
 - Discover interesting patterns in large quantities of data
 - Take decisions based on the patterns



University of Mannheim | IE500 Data Mining | Introduction and Course Outline | Version 1.09.2024 Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."

Data Mining: Definitions



- Idea: mountains of data
 - Where knowledge is mined



Data Mining: Definitions



- Data Mining is a non-trivial process of identifying
 - valid
 - novel
 - potentially useful
 - ultimately understandable
 - patterns in data.

(Fayyad et al. 1996)

- Data Mining methods
 - 1. Detect interesting patterns in large quantities of data
 - 2. Support human decision making by providing such patterns
 - 3. Predict the outcome of a future observation based on the patterns

Origins of Data Mining



- Combines ideas from statistics, machine learning, artificial intelligence, and database systems
- Traditional techniques may be unsuitable due to
 - Large amount of data
 - High dimensionality of data
 - Heterogeneous,
 - distributed nature of data



The Data Mining Process





Source: Fayyad et al. (1996)

Selection and Exploration (1)

- Selection
 - What data is available?
 - What data is potentially useful for the task at hand?
 - What do I know about the quality/provenance of the data?
- Exploration / Profiling
 - Get an initial understanding of the data
 - Calculate basic summarization statistics
 - Visualize the data
 - Identify data problems such as outliers, missing values, duplicate records







Preprocessing and Transformation (2+3)



- Transform data into a representation that is suitable for the chosen data mining methods
 - Number of dimensions (represent relevant information using less attributes)
 - Scales of attributes (nominal, ordinal, numeric)
 - Amount of data (determines hardware requirements)
- Methods
 - Discretization and binarization
 - Feature subset selection / dimensionality reduction
 - Attribute transformation / text to term vector / embeddings
 - Aggregation, sampling
 - Integrate data from multiple sources

Preprocessing and Transformation (2+3)



- Good data preparation is key to producing valid and reliable models
- Data integration/preparation is estimated to take
 70-80% of the time and effort of a data mining project



What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

Advertisement: IE670 Web Data Integration

32

Source: CrowdFlower Data Science Report 2016: http://visit.crowdflower.com/data-science-report.html University of Mannheim | IE500 Data Mining | Introduction and Course Outline | Version 1.09.2024

Data Mining (4)

- Input: Preprocessed Data
- Output: Model / Patterns
- 1. Apply data mining method
- 2. Evaluate resulting model / patterns
- 3. Iterate
 - Experiment with different (hyper-)parameter settings
 - Experiment with multiple alternative methods
 - Improve preprocessing and feature generation
 - Increase amount or quality of training data
 - Combine different methods





Interpretation / Evaluation (5)



- Output of Data Mining
 - Patterns
 - Models
- In the end, we want to derive value from that, e.g.,
 - Gain knowledge



Make better decisions



Increase revenue

Deployment



- Use model in the business context
- Keep iterating in order to maintain and improve model



Tasks and Applications



• **Descriptive** Tasks

- Find patterns in the data
 - E.g. which products are often bought together?
- Predictive Tasks
 - Predict unknown values of a variable
 - Given observations (e.g., from the past)
 - E.g. will a person click a online advertisement?
 - given her browsing history
- Machine Learning Terminology
 - Descriptive = unsupervised
 - Predictive = supervised



Data Mining Tasks



• Classification [Predictive]



• Regression [Predictive]



• Cluster Analysis [Descriptive]



• Association Analysis [Descriptive]





- Previously unseen records should be assigned a class from a given set of classes as accurately as possible.
- Approach:
 - Given a collection of records (**training set**)
 - Each record contains a set of **attributes**
 - One attribute is the **class attribute (label)** that should be predicted
 - Find a model for predicting the class attribute as a function of the values of other attributes

Classification





"tree"



"tree"



"not a tree"



"not a tree"



"tree"



"not a tree"

Classification: Workflow





Classification: Applications



- Credit Risk Assessment
 - Attributes: your age, income, debts, ...
 - Class: are you getting credit by your bank?
- SPAM Detection
 - Attributes: words and header fields of an e-mail
 - Class: regular e-mail or spam e-mail?
- Analysis of tax declaration?
 - Attributes: the values in your tax declaration
 - Class: are you trying to cheat?







Regression

- Predict a value of a continuous variable based on the values of other variables, assuming a linear or nonlinear model
 - Examples:
 - Predicting the price of a house or car
 - Predicting sales amounts of new product based on advertising expenditure
 - Predicting miles per gallon (MPG) of a car as a function of its weight and horsepower
 - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Difference to classification: The predicted attribute is continuous, while classification is used to predict nominal attributes (e.g. yes/no)







Cluster Analysis



- Given a set of data points, each having a set of attributes, and a similarity measure among them, find groups such that
 - Data points in one group are more similar to one another
 - Data points in separate groups are less similar to one another
- Similarity Measures
 - Euclidean distance if attributes are continuous
 - Other task-specific similarity measures
- Goals
 - Intra-cluster distances are minimized
 - Inter-cluster distances are maximized
- Result
 - A descriptive grouping of data points

University of Mannheim | IE500 Data Mining | Introduction and Course Outline | Version 1.09.2024

44

Cluster Analysis: Applications

- Application 1: Market segmentation
 - Find groups of similar customers
 - Where a group may be conceived as a marketing target to be reached with a distinct marketing mix
- Application 2: Document Clustering
 - Find groups of documents that are similar to each other based on terms appearing in them
 - Grouping of articles in Google News









Association Analysis





- Given a set of records each of which contain some number of items from a given collection
- Discover frequent itemsets and produce association rules which will predict occurrence of an item based on occurrences of other items

TID	Items	
1	Bread, Coke, Milk	
2	Beer, Bread	
3	Beer, Coke, Diaper, Milk	
4	Beer, Bread, Diaper, Milk	
5	Coke, Diaper, Milk	

Frequent Itemsets {Diaper, Milk, Beer} {Milk, Coke}

Association Rules {Diaper, Milk} --> {Beer} {Milk} --> {Coke}

Association Analysis: Applications



- Supermarket shelf management
 - To identify items that are bought together by sufficiently many customers



- Process the point-of-sale data collected
 with barcode scanners to find dependencies among items
- Sales Promotion



Frequently Bought Together



Which Methods are Used in Practice?





Classification Algorithms

 itree"
 itree"
 itree"

 itree"
 itree"
 itree"

 itree"
 itree"
 itree"

 itree"
 itree"
 itree"



- Classification:
 - We give the computer a set of labeled examples
 - The computer learns to classify new (unlabeled) examples
- How does that work?
 - K-Nearest-Neighbors
 - Decision Trees
 - Naïve Bayes
 - Support Vector Machines
 - Artificial Neural Networks
 - Deep Neural Networks
 - Many others ...



K-Nearest-Neighbors



- Problem
 - Predict the current weather in a certain place
 - Where there is no weather station
 - How could you do that?
- Symbols
 - Red = Sunny
 - Blue = Cloudy



K-Nearest-Neighbors



- Idea: use the average of the nearest stations
- Example:
 - 2x sunny (red)
 - 3x cloudy (blue)
 - result: cloudy



- This approach is called **K-Nearest-Neighbors**
 - where k is the number of neighbors to consider
 - in the example:
 - k=5
 - "near" denotes geographical proximity

K-Nearest-Neighbor Classifier



Unknown record

- Require three things
 - A set of stored records
 - A distance measure to compute distance between records
 - The value of k, the number of nearest neighbors to consider

K-Nearest-Neighbor Classifier





- To classify an unknown record:
 - Compute distance to each training record
 - Identify k-nearest neighbors
 - Use class labels of nearest
 neighbors to determine the class
 label of unknown record
 - By taking majority vote or
 - By weighing the vote according to distance

Examples of K-Nearest Neighbors



• The k-nearest neighbors of a record x are data points that have the k smallest distances to x



(a) 1-nearest neighbor

(b) 2-nearest neighbor

(c) 3-nearest neighbor

Choosing a Good Value for K



- If k is too small, the result is sensitive to noise points
- If k is too large, the neighborhood may include points from other classes



• Rule of thumb: Test k values between 1 and 20

Discussion of K-NN Classification



• Often very accurate

- for instance for optical character recognition (OCR)
- ... but slow as unseen record needs to be compared to all training examples
- Results depend on choosing a **good proximity measure**
 - attribute weights, asymmetric binary attributes, ...
- KNN can handle decision boundaries which are not parallel to the axes (unlike decision trees)

Decision Boundaries of a k-NN Classifier



- k=1 \bullet
- Single noise points have influence on model



University of Mannheim | IE500 Data Mining | Introduction and Course Outline | Version 1.09.2024

a2

Decision Boundaries of a k-NN Classifier

- k=3
- Boundaries become smoother
- Influence of noise points is reduced



University of Mannheim | IE500 Data Mining | Introduction and Course Outline | Version 1.09.2024



a2

What You Will Learn in This Lecture



- Common data mining tasks
 - How they work
 - When and how to apply them
 - How to interpret their output



Thank you



- Are there any questions?
- Next ...
 - 1. Install the Anaconda Python distribution
 - 2. Get an OpenAl account for using ChatGPT
 - 3. Attend exercise on Thursday
 - 4. Attend the lecture next week