Data Mining Algorithms I (IN5042)

Titel	Data Mining Algorithms I	
Typ	Vorlesung mit Übung	
Credits	6 ECTS	
Lehrform/SWS	3V + 2Ü	
Sprache	Englisch	
Modulniveau	Master	
Arbeitsaufwand	Präsenzstunden	75 Stunden
	Eigenstudium	105 Stunden
	Gesamtaufwand	180 Stunden
Angestrebte Lernergebnisse	Die Studierenden kennen den grundlegenden Prozess des Knowledge Discovery in Datenbanken und die einzelnen Schritte dieses Prozesses sowie grundlegende Problemstellungen im Data Mining (wie z.B. Feature Transformation, Suche nach häufigen Mustern, Musterevaluation). Sie sind in der Lage, Merkmalsräume, Ähnlichkeitsmaße und Distanz-Metriken zu beschreiben, zu analysieren, zu bewerten und gezielt anzuwenden. Sie sind in der Lage, grundlegende Verfahren (wie z.B. Clustering-Methoden, Klassifikatoren, Outlier Detection Methoden, Verfahren des Process Mining) für Anwendungen von Data Mining zu beurteilen, auszuwählen und gezielt einzusetzen sowie die gefundenen Muster und Funktionen zu evaluieren und kritisch zu interpretieren. Sie sind in der Lage, für ein gegebenes Problem (wie z.B. Erkennen von Spam Emails, Warenkorbanalyse, Clustern von Bildinhalten) einen Knowledge Discovery Prozess zu entwerfen und umzusetzen sowie aus den vorhandenen Data Mining Algorithmen geeignete Verfahren auszuwählen und an diese Probleme anzupassen.	
Intended Learning Outcomes	Students are able to reproduce the process of knowledge discovery and fundamental problems of data mining (e.g. feature transformation, frequent pattern mining, pattern evaluation). They are able to describe, to analyze, to evaluate and to apply feature spaces, similarity measures, and distance metrics. They are able to employ and implement methods for data mining tasks (e.g. clustering, classification, outlier detection, and process mining) and to evaluate the resulting patterns and functions. They are able to design and implement knowledge discovery processes	

	for given problem settings (e.g. spam detection, market basket analysis, or image clustering) and to select the best suited data mining methods for these problem settings.
Inhalt	Knowledge Discovery and Data Mining • Definition Knowledge Discovery and Data Mining • KDD Process and its phases • Supervised versus unsupervised learning • Basic tasks of data mining: classification, clustering, outlier detection, frequent pattern mining, process mining. Foundations of Data Mining • Basic characteristics of data • Representation of data, distance functions and similiarity metrics • Presentation of data, data visualization Supervised Learning, Classification • Introduction: task and evaluation of classifiers • Bayesian classifiers, a-priori and a-posteriori probabilities, probability density functions • Linear discriminant functions, dual-class vs multi-class discrimination • Support Vector Machines, maximum margin hyperplane, soft margins • Kernel methods, Kernel SVM • Decision Tree classifiers, pruning techniques • Nearest Neighbor classifiers, instance-based classification, lazy classification • Ensemble classification Unsupervised Learning, Clustering • Introduction, characterization, evaluation of clustering • Partitioning methods (k-means, expectation maximization, k-medoid, etc.) • Density-based and hierarchical clustering (DBSCAN, OPTICS, Single-Link, etc. • Spectral Clustering Outlier Detection • Introduction, task, definitions • Discussion of different methods (e.g., clustering-based, statistical, distance-based, density-based, angle-based, Local Outlier Factor, etc.) Frequent Pattern Mining • Introduction, task, frequent patterns, quality metrics • Apriori principle, anti-monotonicity, pruning

	power • Apriori algorithm, breadth first traversal • Frequent Pattern Tree, depth first traversal • Association rules, market basket analysis • Sequential patterns, frequent subsequences • Interval patterns, overlapping extended objects Process Mining • Introduction, characterization of event-based processes, event-logfiles, tasks, types of process models • Process discovery algorithms (Alpha Miner, Heuristic Miner, Inductive Miner, etc.) • Conformance checking (token replay, alignment), deviation metrics.
Contents	Knowledge Discovery and Data Mining • Definition of Knowledge Discovery and Data Mining • KDD Process and its phases • Supervised versus unsupervised learning • Basic tasks of data mining: classification, clustering, outlier detection, frequent pattern mining, process mining. Foundations of Data Mining • Basic characteristics of data • Representation of data, distance functions and similiarity metrics • Presentation of data, data visualization Supervised Learning, Classification • Introduction: task and evaluation of classifiers • Bayesian classifiers, a-priori and a-posteriori probabilities, probability density functions • Linear discriminant functions, dual-class vs multi-class discrimination • Support Vector Machines, maximum margin hyperplane, soft margins • Kernel methods, Kernel SVM • Decision Tree classifiers, pruning techniques • Nearest Neighbor classifiers, instance-based classification, lazy classification • Ensemble classification Unsupervised Learning, Clustering • Introduction, characterization, evaluation of clustering • Partitioning methods (k-means, expectation maximization, k-medoid, etc.) • Density-based and hierarchical clustering (DBSCAN, OPTICS, Single-Link, etc.

	• Spectral Clustering Outlier Detection • Introduction, task, definitions • Discussion of different methods (e.g., clustering-based, statistical, distance-based, density-based, angle-based, Local Outlier Factor, etc.) Frequent Pattern Mining • Introduction, task, frequent patterns, quality metrics • Apriori principle, anti-monotonicity, pruning power • Apriori algorithm, breadth first traversal • Frequent Pattern Tree, depth first traversal • Association rules, market basket analysis • Sequential patterns, frequent subsequences • Interval patterns, overlapping extended objects Process Mining • Introduction, characterization of event-based processes, event-logfiles, tasks, types of process models • Process discovery algorithms (Alpha Miner, Heuristic Miner, Inductive Miner, etc.) • Conformance checking (token replay, alignment), deviation metrics.
Prüfung	Prüfungsleistung (benotet): -Klausur: 90 min Wiederholungsklausur zu Ende des Semesters. Details werden zu Beginn des Moduls bekannt gegeben. In der Klausur weisen die Studierenden nach, inwieweit sie die vorgestellten Prozesse, Modelle und Verfahren verstanden haben, komprimiert wiedergeben und anwenden sowie auf verwandte Problemstellungen übertragen können. In der Klausur werden 7 bis 10 Aufgaben gestellt, die eine eigenständige Beschreibung der Prozessdefinition, Auswahl und Anwendung grundlegender Verfahren (wie zum Beispiel die Anwendung von Klassifikatoren, Clustering Verfahren und Methoden des Frequent Itemset Mining), Evaluierung von Ergebnissen und Entwurf eines Knowledge Discovery Prozesses erfordern.
Examination	Examination requirements (graded): - written exam: 90 min A makeup exam will be offered at the end of the

	semester, details will be announced at the beginning of the course. Within the written exam, students demonstrate that they understand the presented processes, models, and methods, that they can reproduce and apply them as well as that they can transfer and extend models and methods to similar problems. The written exam consist of 7 to 9 assignments, which require the description of process definitions, selection and application of the basic methods (e.g. the application of classifiers, clustering methods and method for frequent itemset mining etc.), the evaluation of the results and the development of a KDD process for a given task.
Literatur/Literature	• Han J., Kamber M., Pei J. Data Mining: Concepts and Techniques, 3. Auflage, Morgan Kaufmann, 2011 • Tan P.-N., Steinbach M., Kumar V.: Introduction to Data Mining, Addison-Wesley, 2006 • Bishop C. M.: Pattern Recognition and Machine Learning, Springer 2006. • Aggarwal C. C., Reddy C. K. (Eds): Data Clustering, CRC Press 2013. • Mitchell T. M.: Machine Learning, McGraw-Hill, 1997 • Ester M., Sander J.: Knowledge Discovery in Databases: Techniken und Anwendungen, Springer Verlag, 2000 • Witten I. H., Frank E., Hall M. A.: Data Mining: Practical Machine Learning Tools and Techniques 3. Auflage, Morgan Kaufmann, 2011
Medienformen	Folien (Beamer und Handout)
Media	slides show, handouts
Lehr- und Lernmethode	Vorlesung, Übung, Aufgaben zum Selbststudium. Das Modul besteht aus einer Vorlesung und Übungen in kleinen Gruppen (als Tutorübungen). In den Hausaufgaben, die freiwillig abzugeben sind, analysieren die Studierenden die in der Vorlesung vorgestellten Prozesse, Modelle und Verfahren, wenden diese auf konkrete Daten an und erweitern diese für ähnliche Problemstellungen. In den Hausaufgaben werden selbständig anspruchsvolle Übungsaufgaben bearbeitet, die ähnlich zu den Klausuraufgaben sind (siehe oben) und daher zur Vorbereitung darauf dienen. In der Übung werden mögliche Lösungsstrategien der Aufgaben zum Selbststudium diskutiert.
Teaching and Learning	Lecture, tutorial, assignments for individual study.

Methods	Within the assignments (the submission is optional) students analyze the processes, models, and methods presented in the corresponding lectures, apply them to real data, and extend these to similar problems. The assignments consist of demanding problems similar to the assignments in the written exam (for details see above) and serve as a preparation for the exam. Within the tutorials possible approaches for solutions of the assignments will be discussed.
Turnus	Wintersemester
Modulverantwortlicher	Prof. Dr. Thomas Seidl
Dozenten	Prof. Dr. Thomas Seidl