Knowledge Discovery in Datenbanken I (IN5042)

Titel	Knowledge Discovery in Databases I	
Typ	Vorlesung mit Übung	
Credits	6 ECTS	
Lehrform/SWS	3V + 2Ü	
Sprache	Deutsch	
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 die grundliegenden Problemstellung 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, Regressionsverfahren) in verschiedene Bereiche des 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 vorhanden Verfahren 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 (e.g. clustering methods, classification methods, regression methods, outlier detection methods) for data mining tasks and to	

	evaluate the computed 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 und Data Mining • Der KDD Prozess (einzelne Schritte, iterativer Ablauf) • Supervised und Unsupervised Learning • Grundliegende Aufgaben des Data Mining: Klassifikation, Clustering, Outlier Detection, Regression, Frequent Pattern Mining. Merkmalsräume: • Wahrscheinlichkeitsverteilungen (einfache univariate und multivariate Verteilungen, Abhängigkeit von Zufallsvariablen) • Distanzmaße und Ähnlichkeitsmaße (mathematische Eigenschaften wie Reflexivität, Symmetrie, Transitivität) • Beispiele für Feature-Transformationen (z.B. Farbhistogramme, Bag of Words) • einfache Verfahren zu Feature Selection (z.B. greedy forward selection) [optional] • einfache Verfahren der Feature Reduction (z.B. PCA) [optional] Klassifikation: • Evaluation von Klassifikatoren (Testschemata z.B. Crossvalidation, Bootstraping, leaveone-out, Metriken) • Formale Aspekte des Lernens (Generalisierung, Overfitting, Problemdefinition) • Entscheidungsbäume • Bayes-Klassifikationen (naive Bayes, Bayes-Netze, diskrete und kontinuierliche Verteilungen) • Instanzbasierte Klassifikation. • fortgeschrittene Klassifikationsverfahren (z.B. Support Vector Maschinen, Neuronale Netze, Gauss Klassifikatoren, logistische Regression) [optional] • regelbasierte Klassifikation und Inductive logical programming [optional] • Deep Learning Methoden [optional] Regression: • Problemdefinition (Bewertung von

	Regressionsmodellen) • einfache lineare Regressionsmodelle • Grundlegende Verfahren der multivariaten Regression • fortgeschrittene Regressionsverfahren (z.B. kernelbasierte Regression, instanzbasierte Regression). Clustering: • Problemdefinition (Zielsetzung, Abgrenzung zur Klassifikation) • Partitionierende Clusteringmethoden (k-Means, Expectation Maximization, weitere Verfahren z.B. PAM, CLARANCE, k-Modes) • Dichtebasiertes und hierarchische Clustering (z.B. DBSCAN, OPTICS, Single Link) • Self Organizing Maps [optional] • graphbasiertes Clustering und Spectral Clustering [optinaol] • Clusterevaluation [optional] Outlier Detection: • Aufgabenstellung (verschiedene Outlier Definitionen, Abgrenzung zu Clustering und Klassifikation) • statistische Outlier • distanzbasierte Outlierfaktoren • lokale Outlier (z.B. LOF). • fortgeschrittene Verfahren (z.B. ABOD) [optional] • Evaluation von Outlierverfahren [optional] Frequent Itemset Mining und Assoziationsregeln: • Einführung Pattern Mining (Häufigkeit, Konfidenz, Monotonie) • Frequent Itemset Mining (Suchraum, Apriori) • Assozitionsregeln (Ableitung, Interessantheit) • weiterführende Algorithmen zur Berechnung von frequent Itemsets [optional] • Datenstrukturen zur Suche in frequent Itemsets [optional]
Contents	Knowledge Discovery and Data Mining: • Definition Knowledge Discovery and Data Mining • KDD Process (different steps, iterative approach) • Supervised and unsupervised learning • Basic Data Mining tasks: Classification, Clustering, Outlier Detection, Regression, Frequent Pattern Mining. Feature Spaces:

	• Probability distributions (simple univariate and multivariate distributions, dependency of random variables) • Distance and similarity measures (mathematical characteristics such as reflexivity, symmetry, transitivity) • Examples for simple feature transformations (e.g. color histograms, bag of words) • simple methods for feature selection (e.g. greedy forward selection) [optional] • simple methods for feature reduction (e.g. PCA) [optional] Classification: • Classifier evaluation (testing schemes e.g. cross validation, bootstrapping, leave-one-out, evaluation metrics) • Formal aspects of learning (generalization, overfitting) • Decision trees • Bayes classifier (naive Bayes, Bayesian networks) • instance based Classification • advanced classification methods (e.g. support vector machines, neuronal Networks, Gaussian classifiers, logistic regression) [optional] • rule-based classifiers and inductive logical programming [optional] • deep learning [optional] Regression: • Problem definition (Evaluation of regression functions) • Simple linear regression • Basic methods for multivariate regression • Advanced regression methods (e.g. kernel based regression, instance-based regression) Clustering: • Problem definition (aims, difference to classification) • Partitioning clustering methods (k-Means, expectation maximization, further methods e.g. PAM, CLARANCE, k-Modes) • Density-based and hierarchical clustering (e.g. DBSCAN, OPTICS, Single Link) • Self organizing maps [optional] • Graph-based clustering and spectral clustering [optional] • Evaluation of clusterings Outlier Detection:

	• General setting (various outlier definitions, differences to clustering and classification) • statistic outliers • distance-based outliers • local outlier (e.g. LOF) • Advanced methods for outlier detection (e.g. ABOD) [optional] • Evaluation of outlier detection methods [optional] Frequent Itemset Mining and Association Rules: • Introduction to Pattern Mining (Frequency, Confidence, Monotony) • Frequent Itemset Mining (Search space, apriori method) • Association rules (computation, interestingness) • Advanced algorithms for frequent itemset computation [optional] • Data structures to facilitate frequent itemset mining
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 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	Beamer-Präsentation, Tafelpräsentation, Handout
Media	slides show, blackboard presentation, 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 Methods	Lecture, tutorial, assignments for individual study. 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	Sommersemester
Modulverantwortlicher	PD Dr. Matthias Schubert

Dozenten	PD Dr. Matthias Schubert