Syllabus Data mining

1. Introduction to Data Mining
	* What is data mining?
	* Related technologies - Machine Learning, DBMS, OLAP, Statistics
	* Data Mining Goals
	* Stages of the  Data Mining Process
	* Data Mining Techniques
	* Knowledge Representation Methods
	* Applications
	* Example: weather data
2. Data Warehouse and OLAP
	* Data Warehouse and DBMS
	* Multidimensional data model
	* OLAP operations
	* Example: loan data set
3. Data preprocessing
	* Data cleaning
	* Data transformation
	* Data reduction
	* Discretization and generating concept hierarchies
	* Installing Weka 3 Data Mining System
	* Experiments with Weka - filters, discretization
4. Data mining knowledge representation
	* Task relevant data
	* Background knowledge
	* Interestingness measures
	* Representing input data and output knowledge
	* Visualization techniques
	* Experiments with Weka - visualization
5. Attribute-oriented analysis
	* Attribute generalization
	* Attribute relevance
	* Class comparison
	* Statistical measures
	* Experiments with Weka - using filters and statistics
6. Data mining algorithms: Association rules
	* Motivation and terminology
	* Example: mining weather data
	* Basic idea: item sets
	* Generating item sets and rules efficiently
	* Correlation analysis
	* Experiments with Weka - mining association rules
7. Data mining algorithms: Classification
	* Basic learning/mining tasks
	* Inferring rudimentary rules: 1R algorithm
	* Decision trees
	* Covering rules
	* Experiments with Weka - decision trees, rules
8. Data mining algorithms: Prediction
	* The prediction task
	* Statistical (Bayesian) classification
	* Bayesian networks
	* Instance-based methods (nearest neighbor)
	* Linear models
	* Experiments with Weka - Prediction
9. Evaluating what's been learned
	* Basic issues
	* Training and testing
	* Estimating classifier accuracy (holdout, cross-validation, leave-one-out)
	* Combining multiple models (bagging, boosting, stacking)
	* Minimum Description Length Principle (MLD)
	* Experiments with Weka - training and testing
10. Mining real data
	* Preprocessing data from a real medical domain (310 patients with Hepatitis C).
	* Applying various data mining techniques to create a comprehensive and accurate model of the data.
11. Clustering
	* Basic issues in clustering
	* First conceptual clustering system: Cluster/2
	* Partitioning methods: k-means, expectation maximization (EM)
	* Hierarchical methods: distance-based agglomerative and divisible clustering
	* Conceptual clustering: Cobweb
	* Experiments with Weka - k-means, EM, Cobweb
12. Advanced techniques, Data Mining software and applications
	* Text mining: extracting attributes (keywords), structural approaches (parsing, soft parsing).
	* Bayesian approach to classifying text
	* Web mining: classifying web pages, extracting knowledge from the web
	* Data Mining software and applications

**Hodnotenie a práca počas semestra**

Zápočet: Počas semestra budú 2 testy. 40% z celkového hodnotenia

Skúška: Projekt v elektronickej forme. 60% z celkového hodnotenia