ACM Data Science Task Force Course Example

Data Mining
Harbin Institute of Technology, Harbin
Shengfei, Shi

Knowledge Areas that contain competencies (knowledge, skills, and dispositions) covered in the course

Knowledge Area	**Total Number of Contact Hours**
Data mining overview	2
Working with various types of data	6
Knowledge representation and reasoning – probability based	4
Knowledge representation and reasoning – logic based	2
Classification and regression	6
Supervised learning	8
Cluster analysis	14
Outlier detection	2
Pattern mining	4

Where does the course fit in your undergraduate Data Science curriculum?
Data mining is an emerging interdisciplinary subject, covering technologies such as machine learning, statistics, pattern recognition, and artificial intelligence. The purpose of this course is to enable students to fully and in-depth grasp the basic concepts and principles of data mining, master commonly used data mining algorithms, understand the latest developments in data mining, cutting-edge data mining research fields, and data mining technology in different disciplines in the application.

Is this course from or used in other curricula/majors?

What is covered in the course?

Course objective 1: Have the literacy of using mathematical knowledge to solve complex data mining problems, and be able to analyze, design and evaluate various data mining algorithms; have a strong curiosity for new application areas;
Course objective 2: Understand the problem-solving thinking model from engineering problems, to modeling, and then to data mining algorithm design; have the ability to apply data mining algorithms to specific projects;
Curriculum Objective 3: Have a strong ability to pay close attention to and independently learn the latest research results in the field of data mining; be able to analyze the problems and deficiencies of existing research results, and be able to put forward their own independent opinions.

What is the format of the course?
The total number of hours in this course is 48 hours, of which 32 hours are taught in the large class and 16 hours are in the experiment. This is a course that combines theory and application. Classwork provides much of the content and expectations of the course. Curriculum related experiments are designed to allow students to better experience the practical application of the theories they have learned. The teaching of this course will mainly adopt classroom teaching methods, supplemented by the practical links of experimental courses. By guiding students to implement the algorithms taught in the classroom, learn to compare the performance differences of various algorithms, and stimulate students' interest in research and innovation.

How are students assessed?
This course has a score of 100. It consists of:
(1) Homework 20%: evaluated by the number of times and quality of homework completed
(2) Experiment 20%: Determined according to the number and quality of the algorithm
(3) Exam 60%: closed book exam

Course tools and materials
This course does not require any materials, the teacher will send the materials to everyone when teaching. In addition, the following two textbooks can be used as teaching reference for students to supplement reading:
1. "Data Mining: Concepts and Techniques" Author: Jiawei Han (plus) Machinery Industry Press
2. "Introduction to Data Mining" Author: Tan Pang-Ning (US) Posts & Telecom Press

Why do you teach the course this way?
This course is mainly in the form of classroom face-to-face, interspersed with small class discussions, flip classes, and experiments. Classroom face-to-face teaching is mainly taught by teachers, as the most common form of teaching, can be more comprehensive and systematic transfer of the main knowledge points to everyone. Doing experiments can enhance the practical ability to operate, put the knowledge learned into use, not only on paper, general talk, the understanding of knowledge points more in place, more thorough.

Body of Knowledge coverage

KA	Sub-domain	Competencies Covered	Hours
DG	Working with various types of data	Type of data Statistical characteristics of the data Data preprocessing Treatment of missing values	8

		data visualization Program various data preprocessing algorithms and basic data visualization methods.	
AI	Knowledge representation and reasoning – probability based	1 Decision tree induction 1.1 ID3 algorithm 1.2 C4.5 algorithm 1.3 Extract rules from the decision tree 1.4 Overfitting of decision trees 1.5 Decision tree pruning and optimization 1.6 Random Forest Algorithm	4
AI	Knowledge representation and reasoning – logic based	Case-based reasoning	2
DM	Classification and regression	1 Evaluation of the classifier algorithm 2 Programming to implement algorithm.	6
ML	Supervised learning	1 Bayesian classifier 1.1 Bayes' theorem 1.2 Naive Bayes classifier 1.3 Bayesian belief network 2 Support Vector Machine 2.1 The basic idea of support vector machine 2.2 Theoretical basis of support vector machines 2.3 Support Vector Machine Application 3 Case-based learning algorithm 3.2 Partially weighted regression 3.3 Case-based reasoning 4 Regression analysis	8

		4.1 Linear regression 4.2 Logistic regression 5 K-NN classifier	
DM	Cluster analysis	1 Similarity and dissimilarity measures 2 Classification of clustering algorithms 3 Clustering algorithm based on partition 4 hierarchical clustering 5 Density-based clustering algorithm 6 Scalable clustering algorithm 7 Evaluation of cluster quality 8 Programming to realize K-Means, Bisecting K-means, and Dbscan algorithms.	14
DM	Outlier detection	1 Statistical methods 2 Cluster-based detection technology	2
DM	Pattern mining	1 Generate frequent itemsets based on candidate set generation-testing method 2 FP-growth: frequent itemset generation algorithm based on depth first search 3 Evaluation method of association rules	4

Additional topics

Data mining overview. associate analysis

Other comments