Data Mining, CS 565

Course Description:

Data mining is the process of automatically discovering useful information from large data sets or databases. This course will provide an introduction to the main topics and algorithms in data mining and knowledge discovery, including: association discovery, classification, clustering, outlier detection, database support, and so on. Emphasis will be placed on the algorithmic and systems issues, as well as application of mining in real-world problems.

Prerequisites:

Working knowledge of programming and data structures (CS 112, or equivalent). Familiarity with linear algebra, probability and statistics.

Required Text:

Jiawei Han and Micheline Kamber: Data Mining: Concepts and Techniques. Second Edition. Morgan Kaufmann Publishers, March 2006.

Grading (subject to change):

Three programming projects: 35%
Three problem sets: 15 %
Mid-term: 20%
Final exam: 30%
Late penalty: -10% per day, up to three days late. After that no credit will be given.

Tests:

Mid-term, October 29, 2007 in class.
Final, TBA.

Academic Honesty:

Course participants must adhere to the CAS Academic Conduct Code. Copies of the code are available from CAS 105. All instances of academic dishonesty will be reported to the academic conduct committee.

Schedule (subject to change):

Date	Topic	Assignments Out
Sep		
5	Introduction	
10	Data Warehousing: Basic Concepts	
12	Data Warehousing: Design and Implementation	
17	OLAP Evaluation Techniques	Problem Set 1
19	Data Preprocessing and Data Cleaning	
24	Mining Association Rules Algorithms: Definitions and Apriori Algorithm	
26	Algorithms for Frequent Itemset Mining: FP-Tree, MaxMiner	
Oct		
1	Closed and Maximal Itemset Mining : MAFIA, CLOSET	Program 1: Association Rules Mining
3	Association Mining and Correlation Analysis	
9	Scalable Classification Algorithms: SPRINT and SLIQ	
10	Scalable Classification Algorithms: RainForest Framework	
15	Clustering Large Datasets: CLARANS and BIRCH	Problem Set 2
17	Clustering Large Datasets: CURE and DBSCAN	
22	Clustering Large Datasets: CHAMELEON and Graph based Clustering	
24	Projective and Subspace Clustering: CLIQUE and DOC	
29	Mid-Term	
31	Time Series Clustering and Classification	Program 2: Clustering
Nov		
5	Data Streams Mining	
7	Sequential Pattern Mining: SPADE, SPAM	
14	Sequential Pattern Mining: Mining Periodic Patterns	
19	Sequential Pattern Mining with Constraints	
21	Graph Mining	Program 3: Sequential Mining
26	Privacy Preserving Data Mining	
28	Privacy Preserving Data Mining (cont.)	
Dec		
3	Outlier Detection: Basic Concepts and Definitions	
5	Outlier Detection Algorithms	Problem Set 3
10	Spatial Data Mining	
12	Spatio-temporal Data Mining	