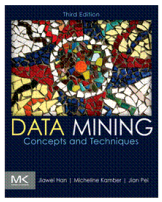
CS 521: DATA MINING TECHNIQUES

Description: This course covers data mining topics from basic to advanced level. Topics include data cleaning, clustering, classification, outlier detection, association-rule discovery, tools and technologies for data mining and algorithms for mining complex data such as graphs, text and sequences. Students will work on a data mining project to gather hands-on experience.

The course learning objectives include

- Learning basic data mining algorithms and their applications
- Learning about the tools and technologies available for analyzing various types of data
- Gaining hands-on experience in cleaning, managing and processing complex data.

Book: Data Mining: Concepts and Techniques, 3rd ed. By Jiawei Han, Micheline Kamber and Jian Pei, The Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, July 2011. ISBN 978-0123814791. We will be occasionally referring to this book by Charu Aggarwal. The book is freely available to download in campus network.



Grading: Grading: There will a final exam worth 35% of the grade. Students will pick datasets for projects and apply mining algorithms. Project is worth 40%. There will be three to five homework, together they are 20% of the course. Homework will focus on understanding the algorithms and techniques. Remaining 5% will be on class participation and attendance.

Lecture Schedule: A tentative weekly distribution of topics is given below. There will be re-arrangement for holidays and exams.

Week 1:	Ch. 1, 2: What is Data Mining? Types of Data.
Week 2:	Ch. 3: Data Preprocessing. Cleaning, Integration, Reduction and Transformation
Week 3:	Ch. 6, 7: Mining Frequent Patterns (FP), Associations and Correlations. Apriori, FP Tree
Week 4:	Ch. 8: Basic Classification: Decision Tree, Bayes Classifier, Rule Based, Goodness measures
Week 5:	Ch. 8, 9: Advanced Classification: Boosting, Bagging, Random Forest, Lazy Learners, FP based classification
Week 6:	Ch. 10: Basic Clustering: Hierarchical, Partitioning, Density-based, Grid-based

Week 7:	Ch. 11: Advanced Clustering: Subspace clustering, Co-clustering, Fuzzy clustering, Expectation-Maximization clustering
Week 8:	Ch. 12: Outlier Detection: Statistical and Proximity based methods
Week 9:	Ch. 13: Mining Complex Data Types: Sequences (real and discrete)
Week 10	Mining Complex Data Types: Graphs and Trees
Week 11:	Mining Complex Data Types: Text, Logs, Reviews
Week 12:	Ch. 4,5: Data Mining Systems: Data warehousing, Data cubing, Business Intelligence systems
Week 13:	Data Mining Tools: Weka, Vowpal-wabbit, Pivot-tables, Matlab Statistics Toolbox
Week 14:	Web Mining: Web search, Computational advertising, User behavior modeling, Fraud detection

Project: Each student will do one project. A project consists of four phases with *equal weights*.

1. Classification: Perform classification on the chosen dataset and produce cross-validated precision/recall numbers.
2. Clustering: Perform clustering on the chosen dataset and produce meaningful clusters.
3. Outlier Detection: Perform outlier detection algorithms on the given dataset and identify anomalous behavior.
4. Ensembling: Perform an ensembling technique to improve accuracy of any of the above tasks.

In each phase, a student produces a report describing data cleaning, method(s), results, and discussions. Phase specific goals will be announced in the class page. A student will merge four small reports in a final report and submit in the finals week.

Exam:

The exam will cover everything taught in the class. Questions will be deterministically testing the student's knowledge about the algorithms. Two sample questions are attached.

1.

Age	Graduated	Income	Gender	Smokes?
18	No	12K	M	Yes
22	Yes	32K	F	Yes
27	No	59K	M	Yes
10	No	0K	F	No
29	Yes	65K	F	No
45	No	47K	F	No
59	Yes	75K	M	No
37	Yes	72K	M	Yes
31	No	32K	M	No

a. What are the information gains for the attributes Graduated and Gender? You don't have to compute the log functions. Analytical expression with numbers are fine.

The formula for information gain of an attribute is the following.

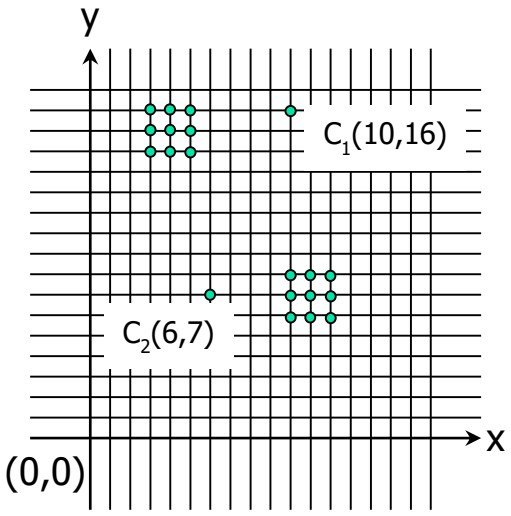
InfoGain = Info(D) – $\text{Info}_A(D)$ where

$\text{Info}(D) = -\sum_{i=1\ to\ m} p_i \log(p_i)$, m is the number of classes and

$\text{Info}_A(D) = \sum_{j=1\ to\ v}(|D_j|/|D|)\ \text{Info}(D_j)$, v is the number of values of attribute A.

b. Draw any decision tree that fits the above data.

2.



Assume C_1 and C_2 are the two initial cluster centers. Assume each grid is one unit and K = 2. Find the new cluster centers for the next iteration if you are clustering using the K-means algorithm.