COSC6335: Data Mining

Instructor: Carlos Ordonez

Course information

Schedule: Tu Th 11:30-13:00
Office hours: Tu Th 10:30-11:30
email: ordonez@cs (Start subject line with "COSC6335-")
TA: Zhibo Chen (zchen6@cs.uh.edu)

1 Course contents

This is a graduate level database course on data mining. The focus of the course is understanding how machine learning and statistical techniques work in a database system, especially for large data sets. The course has some overlap with machine learning, but it is less theoretical and more systems oriented. Notice the course is self-contained: it will cover the basics of relational DBMSs and SQL. The textbook is [2], complemented by [1, 3]. The course will also require reading some research papers.

Topics include the following. Introduction to Data Mining: definitions, history, modern applications. Data Mining process: stages, iterative. Overview of a relational DBMS: architecture, SQL. Statistical analysis on one variable: mean, variance, mode, quantiles, histograms. Data quality: nulls, skewed distributions, referential integrity. Data preprocessing: denormalization, coding, scaling, null handling, aggregations. OLAP and cubes: fact and dimension tables, roll-up, drill down, pivoting. Association rules: metrics, data structures, extensions such as sequences, graphs. Clustering: hierarchical, spatial, K-means, EM. Dimensionality reduction: feature selection, PCA, Factor Analysis. Classification: Bayesian classification, nearest neighbor, decision trees, SVMs. Regression: linear, logistic. Scoring: clustrering, regression, decision trees, Bayesian classification. Advanced topics: streams, UDFs, time series, Bayesian statistics, non-linear regression, neural nets.

2 Grading

- 70%: 8 assignments.
- 30%: Final exam.

All homeworks must be turned in to get B-. The course will use a relational DBMS and a modern data mining tool. Assigments will involve programming, using a data mining tool and writing some reports. Written reports are preferred in Latex/PDF, but any word processor is acceptable. Programs will be developed in Java and SQL (C++ or C# are not allowed). We will use scientific data sets for analysis. All assignments must be done in pairs (team of 2 students).

References

[1] R. Elmasri and S. B. Navathe. *Fundamentals of Database Systems*. Addison/Wesley, Redwood City, California, 3rd edition, 2000.

[2] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 1st edition, 2001.

[3] T.M. Mitchell. *Machine Learning*. Mac-Graw Hill, New York, 1997.