CSC 468:Knowledge Discovery from Data (KDD)
Spring 2009
Course Syllabus

March 30, 2009

Instructor: Alexander Dekhtyar
email: dekhtyar@csc.calpoly.edu
office: 14-215

What	When		Where
Lecture	TR	3:10 – 4:30pm	186-C203
Lab	TR	4:40 – 6:00pm	14-302
Final Exam	June 9, 2009 (Tuesday)	4:10 - 7:00pm	186-C203

Note: the class will not have a written final exam, but I reserve the right to use the final exam meeting time for class-related activities held in lieu of the exam.

Office Hours

	When	Where
Tuesday	11:10am - 12:00pm	14-215
Wednsday	9:10am - 12:00pm	14-215
Thursday	11:10am - 12:00pm	14-215

Additional appointments can be scheduled by emailing the instructor at *dekhtyar@csc.calpoly.edu*.

Description

This class is intended as an overview of a wide range of modern intelligent information analysis fields and technologies. The course is intended for senior students in Computer Science, Software Engineering and Computer Engineering majors. The course give a broad overview of OLAP (On-Line Analytical Processing) and data warehousing, data mining (association rules mining, classification, clustering), information filtering and recommender systems, information retrieval and web search, and web mining.

Learning Objectives

After taking the course the students are expected to be able to

1. **recognize** different types of KDD procedures and **identify** their uses;
2. **implement** algorithms/methods/techniques for KDD tasks to **solve** KDD problems;
3. **interpret** and **analyze** the *results* of KDD processes;
4. **recognize** and **evaluate** *societal impact* of KDD technology, **make informed choices** about use of KDD technology.

Texbook

- Bing Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, Springer, 1st ed. 2007. ISBN: 978-3540378815.

This book is **mandatory**. It contains all material studied in the course (and then some) except for the Data Warehousing/OLAP and Collaborative Filtering parts of the course. The Data Warehousing/OLAP part of the course will come from

- A. Silberschatz, H.F. Korth, S. Sudarshan. *Database System Concepts*, 5th Ed., McGraw Hill, 2006. ISBN: 0-07-295886-3. (Chapter 18)

Collaborative Filtering part of the course will be covered using original papers on the topic and the instructor's lecture notes.

Topics

No.	Topic	Duration (weeks) (weeks)	Material
1.	Data Warehousing, OLAP	1	Chapter 18 (Silberschatz)
2.	Association Rules	1	Chapter 2 (Liu)
3.	Supervised Learning (Classification)	1.5	Chapter 3 (Liu)
4.	Unsupervised Learning (Clustering)	1.5	Chapter 4 (Liu)
5.	Collaborative Filtering	1	*notes*
6.	Information Retrieval	2	Chapter 6 (Liu)
7.	Link Analysis	1	Chapter 7.3 (Liu)

Grading

Labs and Homeworks	60-70%
KDD assignment	15-20%
Design assignment	15-20%

Course Policies

Prerequisites

This course has two prerequisites and a *recommended* course. The prerequistes are CSC/CPE 365, Intro to Databases and STAT 350/STAT 312 or any other equivalent statistics course. It is also recommended that you have taken CSC 480, Artificial Intelligence.

In this course, I will assume that you are familiar with the main concepts studied in an undergraduate database course. In particular, I will assume familiarity with ***Relational Database model*** and **SQL**.

The first lab of the course is used to test your knowledge of SQL, and your ability to work with Oracle DBMS using sql*plus environment. At least one other assignment in the course involves direct work with Oracle DBMS, and you may find it useful to use a DBMS for data storage/processing in some later assignments.

Exams

The course will have **no exams**. In their stead you will be offered two take-home group assignments. One assignment will ask you to apply the methods studied in the course to exploratory analysis of some real-life data. The other assignment will ask you to present a system design for a software system/business process that involves, as a major part, a KDD component. Both assignments will be given out around the mid-point of the course and will be due at the end of the course.

We **may use the reserved final examination time** for activities associated with either or both assignments.

Labs, Homeworks

Hands-on KDD-related activities are the major part of the course. Some activities will be set up as lab exercises, some other activities may be offered as purely take-home assignments (this will be determined by the pace of the course).

Each lab/homework assignment will involve some data analysis task, that may involve using existing software, software provided by the instructor as well as (and mostly) the software developed by you. The course concentrates on **basic algorithms for performing standard KDD tasks**: the labs/assignments give you an opportunity to cement the knoweldge of the algorithms covered in class.

We will have pair programming, small team and individual assignments. Each lab/assignment will specify what type it is and how the pairs/teams are formed.

Late Submissions

*Late **lab** and **assignment** submissions* are strongly discouraged. The course will run on a tight schedule, and not submitting on time will lead to time carved out of the next assignment. A penalty of 10 - 30% will be assessed for any submissions that are late by less than 24 hours. No credit will be given for any later submissions. You are encouraged to submit on time even if your submission it is not perfect. You can then resubmit a fixed version late, subject to the abovementioned rules. When more than one submission is present, we will independently grade two submissions: (i) the latest on-time submission and (ii) the latest late submission for which non-zero credit can be assessed. Your grade for the project will be the **maximum** of the two grades.

Communication

The class will have an official mailing list. The email address for the mailing list is *csc-466-01-2094@calpoly.edu*. All students enrolled in the class are automatically subscribed to the mailing list (using the email addresss that the CS department has on file).

I encourage questions during classtime and questions via email. My answers to email questions may be broadcast to the entire class via the mailing list, if the answer may be relevant to everyone (e.g. a correction in a text of a handout, or a clarification of a homework problem), and may also appear on the web page. The questions can also be posted to the mailing list directly. The mailing list will also be used for all annoucements related to the course. It is your responsibility to read your class-related email. Failure to read email posted to the mailing list cannot be used as an excuse in the class.

Web Page

Class web page can be found at

http://www.csc.calpoly.edu/~dekhtyar/466-Spring2009

Through this page you will be able to access all class handouts including homeworks, lab assignments, project information, lab/project data and lecture notes.

Links to additional information, and notes and announcements will also be posted.

Wikis

The course has its own wiki page at

http://wiki.csc.calpoly.edu/csc466-2009

All students will have read/write permissions on the wiki. The wiki will be used as the workspace for individual and team assignments (both for some of the regular lab assignments and for the two big assignments).

Some of the data you will be using in the course will be posted to the datasets wiki:

http://wiki.csc.calpoly.edu/datasets

You will have read-only access to the parts of the wiki relevant to the course.

Academic Integrity

University Policies

Cal Poly's Academic Integrity policies are found at

http://www.academicprograms.calpoly.edu/academicpolicies/Cheating.htm

In particular, these policies define *cheating* as (684.1)

> "*...obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another's test or examination; discussion of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, "cheat sheets," or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same.*"

Plagiarism, per University policies is defined as (684.3)

> "... the act of using the ideas or work of another person or persons as if they were one's own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary."

University policies state (684.2): "Cheating requires an "F" course grade and further attendance in the course is prohibited." (appeal process is also outlined, see the web site above for details.). Plagiarism, per university policies (684.4) can be treated as a form of cheating, although a level of discretion is given to the instructor, allowing the instuctor to determine the causes of plagiarism and effect other means of remedy. It is the obligation of the instructor to inform the student that a penalty is being assessed in such cases.

Course Policies

All homeworks are to be completed by each student **individually**. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above (e.g., you can solve together a problem from the textbook that had not been assigned in the homework, but you should solve assigned problems individually).