CSCI 48100 Section 26607, Data Mining

Spring 2017 Syllabus

Class Location and Time:	SL 051, Monday, Wednesday at 10:30 am - 11:45 am
Instructor:	Mohammad Al Hasan (office: SL 277, phone: 274-3862, email: alhasan@iupui.edu)
Office Hour:	Tuesday 2:30 pm - 3:30 pm and by appointment
Piazza Course Webpage:	`piazza.com/iupui/spring2017/cs481/home`
TA:	Vachik Dave (vsdave@iupui.edu)
TA office Hour and Location:	TBA

Course Description

The topics covered in this course can be decomposed in the following five parts: i) data analysis, ii) frequent pattern mining, iii) classification methods, iv) clustering methods, and v) applications and trends in data mining. We may cover some topics that are not directly tied to any of the above five parts or that may be tied to more than one parts. In data analysis, we will mainly discuss background materials from linear algebra and probability that are essential for effective data representation. For instance, linear algebra based methods, such as, PCA (Principal Component Analysis) and SVD (Singular Value Decomposition) will be covered. In pattern mining, our focus will be to study the *frequent pattern mining* algorithms for mining patterns from set, and sequence datasets. In classification, we will cover some of the supervised classification methods (kNN, neural networks, Bayesian, and SVM), including methods for evaluating classification systems. In clustering, we will cover some clustering methods, such as k-Means, hierarchical, and density-based clustering. Finally, if time permits, we will discuss some interesting applications of data mining methods in e-commerce, and social networks.

Textbook

- (**ZW**) Data Mining and Analysis: Foundations and Algorithms, by Zaki, and Wagner, Cambridge University Press, 2014. This book is available to download for free from `http://www.dataminingbook.info/uploads/book.pdf`.
- (**TSK**) Introduction to Data Mining, by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (1st edition). This book is optional. There are three free chapters that you can download from the book home page (`http://www-users.cs.umn.edu/~kumar/dmbook/index.php`)

Pre-requisite

You should have some background on statistics, and linear algebra. You also need to know how to program in C++/Java, Matlab/R, and Python.

Grading Criteria

There are total 100 points: homeworks (30 points), first midterm exam (15 points), second midterm exam (15 points), final exam (30 points), in-class discussion and piazza activities (10 points). Final letter grade will be determined based on the following table*.

Grade Category	**Grade**	**Score Required**	**Relative Class Rank Range**
Exceptional	A+	> 93	Top 10%
Outstanding	A	$[86, 93)$	Top 20%
Very Good	A-	$[80, 86)$	Top 30%
Good	B+	$[75, 80)$	Median%
Satisfactory	B	$[70, 75)$	Lower 40%
Reasonable (Pass)	B-	$[65, 70)$	Lower 25%
Marginal (Pass)	C+, C, C-	$[50, 65)$	Lower 10%
Fail	F	$[0, 50)$	-%

* Actual grade will be determined using the above table as reference by balancing the *score required* and *relative class rank* factors.

Homeworks

Homework assignments are generally based on implementation of a data mining method using one of the high level languages. For solving the homework questions, you may need to write small Matlab/R/Python scripts. The instructors prefer Python (specifically, the Numpy package of Python) over Matlab/R and hence, the students who use Python will receive active help from the TA, in terms of installing or locating the necessary libraries for solving their assignments. However, the students are allowed to use Matlab/R as an alternative of Python, if they are already comfortable with these languages.

There will be 3 to 4 homeworks. Homeworks are typically due in two weeks. Over the entire semester, there are 5 late days that you can use at your convenience. Once, you have used up all the 5 days, no late homeworks will be accepted.

Topics (Tentative)

Below we show the topics that we plan to cover in this class.

- **Data Analysis Foundation: 3 weeks** (Algebraic and probabilistic view, Numerical Attributes, Categorical Attributes, Covariance Matrix, Dimensionality Reduction)
- **Clustering: 3 weeks** (K-means, Hierarchical, Density-based, Graph Clustering)
- **Classification: 3 weeks** (Learning theory, Decision tree, Bayesian, k-NN, Neural Networks, classifier assessment, challenges)
- **Pattern mining: 2 weeks** (Itemset mining, Sequence mining, Network analysis)
- **Data Mining applications: 3 weeks** (Text Mining, Feature Selection, Recommendation Systems)

Academic Integrity and Collaboration Policy

Homework assignments are supposed to be individual work. You can discuss homework problems with your classmates, but the submitted assignment must be entirely your own work. Software will be used to identify code similarity. If evidences of collaboration are found, all the engaging parties will receive zero point for that entire assignment, and the incident will be reported to the head of the department. For successive involvement, the student will be given a letter grade F for that course and the incident with be reported to the appropriate authorities of IUPUI. Please see IU Code of student Ethics (`http://www.hoosiers.iupui.edu/studcode/stucode.htm`) and ACM's code of Ethics and Professional Conduct (`http://www.acm.org/constitution/code.html`) for standard of ethical conduct. IUPUI CS department has also drafted an academic dishonesty policy document, which is available at the end of this document. It is your responsibility that you read and understand the policy.

The `piazza.com` online forum is being used for providing a collaboration platform that is transparent, and can be monitored by the instructors. Students are encouraged to use this forum extensively and they will be awarded points based on their activities on this forum. If you need clarification on a topic that is taught in the class, or you need a hint to solve a homework question, instead of asking your fellow students offline, post your questions on piazza forum. Instructors or TAs will answer the questions. Other students are also welcomed to answer questions posted on the forum. However, be careful not to post a complete (or a significant part of the) solution.

Important Dates

- **January 15, 2017**: Last day to Withdraw with 100% refund.

- **January 16, 2017**: Withdrawal with automatic grade of W begins
- **Feb 20, 2017 (in class)**: First Midterm Examination
- **March 12, 2017**: Last day to Withdraw with automatic grade of **W**.
- **March 13,2017 - March 19, 2017**: Spring Break
- **April 03, 2017 (in class)**: Second Midterm Examination
- **May 05, 2017 (10:30 am - 12:30 pm)**: Final Examination

Policy on Academic Dishonesty in the Department of Computer and Information Science at IUPUI

The faculty in the Department of Computer and Information Science (henceforth, referred to as the department) values academic honesty to be absolutely essential and expects all students to conform to it. Any violation of academic integrity is considered a serious offense and will result in severe consequences.

The policy against violations of academic integrity will be enforced at the departmental level across all courses.

If a student does not abide by this policy then, for the first violation, he/she will receive zero points for the component of the course on which academic misconduct occurred and will be reported to the Department Chairperson. If the violation is not related to a specific assignment or exam, the course instructor reserves the right to impose the zero-point penalty to any component of the course.

For a second violation of academic integrity (occurring anywhere in the graduate or undergraduate curriculum, in the same or a different semester, in the same or a different course), the student will receive a failing grade for the course where the second violation occurred, as enforced by the Department Chair and the School of Science Deans Office, and, in addition, an official reporting process will be initiated by the Department Chair as per IUPUI's Student Conduct Policies: http://studentaffairs.iupui.edu/student-rights/student-code/.

For a third violation, the department will initiate dismissal request from the program in which the student is enrolled.

In all cases of academic integrity violation, the involved student will be notified in writing at the time the offense is observed and acknowledge the receipt of such notice in writing.

This is the minimal policy and the department reserves the right to impose more severe penalties for the first and/or second offense of academic misconduct.

The student will have opportunities to file appeals at the department, the school, and the university levels, to contest the academic dishonesty finding and/or the imposed penalty.

At the department level, any appeal will be made to the departments graduate or undergraduate committees respectively, depending on whether the student is a graduate or an undergraduate student. The graduate or undergraduate committee chair will substitute any committee members involved in the penalty imposition process with other faculty members with no conflicts of interest before processing the appeal. If desired, a student can pursue a further appeal to the School of Science Appeals Committee. Finally, the student can also submit an appeal to the IUPUI Appeals Committee.