AC 2010-1357: HOW A DATA MINING COURSE SHOULD BE TAUGHT IN AN UNDERGRADUATE COMPUTER SCIENCE CURRICULUM

Reza Sanati-Mehrizy, Utah Valley University

Paymon Sanati-Mehrizy, University of Pennsylvania

Afsaneh Minaie, Utah Valley University

Chad Dean, Utah Valley University

How a Data Mining Course Should be Taught in an Undergraduate Computer Science Curriculum

Abstract

Data mining is a relatively new area of computer science that brings the concept of artificial intelligence, data structures, statistics, and database together. It is a high demand area because many organizations and businesses can benefit from it. There is no doubt that it is a great idea to teach a data mining course in computer science curriculum. As you can tell, students taking a data mining course need to have background in quite a few areas to be successful. Not every student taking this course may have the background required in all these areas. The question is how can an instructor remedy the challenge of teaching a group of students with widely-ranging backgrounds, and at what level should this course be taught. Furthermore, the issue of group work arises, specifically as to whether data mining course projects should be accomplished individually or as teams.

Studies show that many universities are teaching data mining course(s) within their computers science curriculum. Each school teaches this course according to their individual methodology. This paper will study different teaching approaches used by different institutions around the country and makes some appropriate teaching recommendations based on the conclusions reached.

Introduction

To research how data mining courses are being taught at the undergraduate level, we searched the websites of computer science departments of various universities around the United States looking for data mining course syllabus, schedule and related supporting materials. The syllabus, schedule and material were then analyzed to discern the basic structure and focus of the class. Not all schools that offer courses in data mining offer them at the undergraduate level. Graduate level courses were not considered for this study but it is interesting to note that graduate level courses follow the same basic patterns discovered in the undergraduate courses.

In this study, we discovered four different teaching models used by different universities to teaching data mining at the undergraduate level. These four different models are: Mathematical/Algorithm Based, Textbook Based, Topical Based, and Applied Data Mining. Of the nine schools we researched, five followed the same pattern. The remaining four schools are divided among three different approaches as showed in figure 1.

	Mathematical / Algorithm Based	Textbook Based	Topic Based	Applied Data Mining
Brigham Young University			x	
Carnegie Mellon	x			
Central Washington University	x			
MIT	x			
Princeton	x			
Purdue		x		
Stanford				x
University of Illinois at Urbana-Champaign		x		
University of Vermont	x			

Figure 1

In the following sections, we will discuss these four teaching models starting with the most popular one, Mathematical / Algorithm Based.

Mathematical / Algorithm Based

Data mining has its roots in statistics. It should not be a surprise that the most popular method for teaching a course is one based on the mathematics of how various algorithms are derived and applied. These courses typically start out with a quick review of basic statistics principles like probability, distribution, the chain rule and Bayes rule. They then focus the remainder of the term discussing specific methods used in data mining. These methods generally include clustering, linear regression, classification trees and distribution. Each method is introduced or derived as a mathematical formula and is explored from a theoretical point of view. The focus of these courses appears to be more on theory and less on practical application. These courses do not appear to follow any specific textbook and tend to use research papers as supporting material.

The data mining course taught at Princeton[2] is a good example of this pattern. The course starts with a quick review of statistics and then covers Naïve Bayes classification, clustering, Markov models, and linear regression. Each topic is covered with an emphasis on the mathematics involved and how the various formulas are derived. The course ends with two guest lecturers who cover how data mining techniques are applied in their field of interest.

The statistics department at Carnegie Mellon[8] offers their course in data mining. The course covers clustering, regression, prediction, classification and distribution. The statistical programming language R is used to program assignments related to the topics covered.

The course at MIT[7] is a mathematics based approach. They also discuss various algorithms and how to implement them but also cover applications of the techniques they are studying including grouping pictures with similar content and classifying video. MIT also uses the programming language R in their lab assignments.

The Central Washington University[6] data mining course introduces basic data mining algorithms and statistical methods during the first week and then divides the semester into four main parts, numerical, rule based, relational, and fuzzy logic data mining. Each section appears to have a corresponding project.

The University of Vermont[9] lists data mining as one of three areas of department academic research. The main areas of focus for their data mining course are decision trees, rule induction, Bayesian networks, association analysis, and sequential patterns. As they cover each of these areas, appropriate algorithms are introduced and discussed. They use the Weka[10] software package to demonstrate some of the principles being discussed. The last third of the course is devoted to the presentation of papers. This course appears to have a heavy mathematical emphasis.

Textbook Based

The textbook approach may have similar qualities to the mathematical / algorithm based courses but is noticeably different in that they are structured around a specific text and follow it almost exclusively. The course will start with chapter one of the text and progresses steadily through the book. This method also focuses on clustering, linear regression, classification trees and distribution but in the same order and manner as the text.

The University of Illinois at Urbana-Champaign[5] is the first school we will look at that follows this approach. The course starts with an introduction to data mining and moves into data preprocessing. Next they study data warehousing and the application of the data cube. The first algorithms to be introduced are those used to find association rules in data. The course next looks at classification and prediction methods. The last topic covered is clustering. This University lists data mining as an area of department research. It is unclear, from the online materials we found, how much the research focus on mining data streams and sequential and semi-structured data that the department is involved with is integrated into their course.

The course offered at Purdue[3] University also follows the textbook approach. Their course starts out with an introduction to the subject followed by data preparation. Association rule is the first algorithm covered in the course. They then move on to classification, prediction and finally clustering. Purdue University finishes the course covering time series mining by looking into how to mine data streams. They end the class with a look at how data mining is applied and used to detect fraud.

Topic Based

The topic based course does not follow a specific textbook nor is it primarily focused on the derivation and intricacies of the various algorithms used in data mining. This style of course mixes the standard topics of clustering, linear regression, classification trees and distribution with real world applications. Instead of using a single textbook, a combination of research papers, web articles and selections from various books are used to provide background for the topic to be covered.

Brigham Young University[4] offers a course that fits this description. It starts with a look at how data mining has been applied in a successful industry to produce actionable results. Next they look at machine learning and the data mining process. The next topic is business understanding, the process of determining the data mining goals and producing plans to achieve them. The course then gets into the standard course topics of association rules, linear regression, clustering and graphs and then looks at how data mining is used on the web, in the field of medicine, and social network sites. The course ends with a discussion of data mining ethics, past data mining blunders, how to plan successful data mining projects, and students' projects presentations.

Applied Data Mining

Applied data mining combines theory with hands-on application. This method is similar to the topical based method in that it does not use a textbook and pulls its supplementary material from slides and handouts created by the instructors. It differs in that it applies the concepts being taught to current real world problems.

The course offered at Stanford[1] is unique among the courses we looked at in this study. Stanford is adjacent to Palo Alto California and approximately 20 miles from San Jose. It is in other words located adjacent to Silicon Valley and has access to some of the largest computer, software and web based companies in the world. Stanford takes advantage of its location and gears its class toward programming the various data mining algorithms and applying them to real world commercial applications like the Netfilx movie recommendation system, Ecommerce shopping cart analysis and search engine web page analysis. Each area of application is backed up with a healthy dose of theory and mathematics.

Recommendations

Data mining is an area of study that combines both Computer Science and Statistics. The material covered is not typically covered in an introductory Statistics class or in other Computer Science courses. Covering this complex but valuable topic can be a challenge at the undergraduate level. We have studied the approaches taken by nine universities and have noted that there are different approaches an instructor could take and every one of them have good reason for taking that particular approach. The objective of this study was not to identify the best data mining course but the intention was to decide which style of teaching would work for our students. Of all the approaches we studied the course we would like to take ourselves does not exist but if it did it would be a hybrid between Stanford and Brigham Young University. We like the topic based approach of BYU and feel that it exposes the students to how data mining is currently being done in industry and gives them a good feel for what is involved and the processes you go through. The apparent hands-on nature of the course at Stanford is also quite intriguing. The reason that we feel this way is because students taking a traditional, Statistics based course may understand the theory but without applying it may not fully appreciate the potential of the applying that theory to solve real world problems and because our students mostly are employed and like to learn by doing the projects. We plan to use the text book by Ian H. Witten and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations", because of its tutorial section of Weka software package.

Undergraduate students taking a Data Mining course may have some areas of weakness in their background that may need to be addressed by the instructor. The instructor should be aware of this and be willing to alter the pace of the course as necessary. It is more important for the students to learn than to keep to a schedule. The instructor should initially schedule some optional material into the course that can be skipped as necessary if students require extra instruction in core areas of the course.

If the instructor plans on having the students complete a final project we recommend that they have the students work in teams. In any team there is the possibility that some members may not pull their own weight but in a group, students can learn from their peers and be exposed to alternate methods and solutions to the problem they are working on. Individually some students will inevitably excel but as a group those who would not excel on their own may work with those who will and hopefully learn from their peers and improve their own performance.

References

[1] Ajaraman, A., & Ullman, J. (2003). CS345A, Winter 2009: Data Mining. Retrieved Nov. 19, 2009, from Stanford University, Stanford, CA. Web site: http://infolab.stanford.edu/~ullman/mining/2009/index.html.

[2] Blei, D. (2003). COS424: Interacting with Data. Retrieved Nov. 19, 2009, from Princeton University, Princeton, New Jersey. Web site: http://www.cs.princeton.edu/courses/archive/spring08/cos424/syllabus.html.

[3] Clifton, C. (2003). CS 490D: Introduction to Data Mining. Retrieved Nov. 19, 2009, from Purdue University, West Lafayette, Indiana. Web site: http://www.cs.purdue.edu/homes/clifton/cs490d/.

[4] Giraud-Carrier, C. (2003). CS 476 Introduction to Data Mining. Retrieved Nov. 19,

2009, from Brigham Young University, Provo, UT. Web site: http://cs.byu.edu/courses/cs476.

[5] Han, J. (2003). CS412 Fall 2008. Introduction to Data Warehousing and Data Mining. Retrieved Nov. 19, 2009, from University of Illinois, Urbana, IL. Web site: http://www.cs.illinois.edu/~hanj/cs412/index.htm.

[6] Kovalerchuk, B. (2003). CS 456: Data Mining. Retrieved Nov. 19, 2009, from Central Washington University, Ellensburg, WA. Web site: http://www.cwu.edu/~borisk/456/syllabus.html.

[7] Minka, T. (2003). 36-350: Data Mining. Retrieved Nov. 19, 2009, from Massachusetts Institute of Technology, Cambridge, MA. Web site: http://alumni.media.mit.edu/~tpminka/courses/36-350/.

[8] Shalizi, C. (2003). Statistics 36-350: Data Mining. Retrieved Nov. 19, 2009, from Carnegie Mellon, Pittsburgh, PA. Web site: http://www.stat.cmu.edu/~cshalizi/350/.

[9] Xindong, W. (2003). CS 331/295: Data Mining. Retrieved Nov. 19, 2009, from University of Vermont, Burlington, VT. Web site: http://www.cs.uvm.edu/~xwu/kdd/.

[10] Weka – Data Mining with Open Source Machine Learning Software, www.cs.waikato.ac.nz/ml/weka.