ADDIS ABABA UNIVERSITY
BUSINESS INTELLIGENCE, DATA WAREHOUSING, AND DATA MINING
CIT 828: Summer 2009

Instructor(s)	**Dr. Shawndra Hill (contact Instructor)**
Course Webpage	https://shawndra.pbwiki.com, send requests to course team lead for login credentials
Classroom	TBD
First/Last Class	July 6, 2009/July 17,2009
Class times	9am-1pm Monday – Friday
Local Instructor	Mr. Sebsibe Hailemariam
Office Hours	1pm – 2:30 pm or by appointment
Email	shawndrahill@gmail.com **subject: [DSS class] … ← note!**
Telephone	TBD, Skype: shawndra
Prerequisites	Admission to the IS PhD Program
Textbook and Resources	Online resources. See class website.
Supporting Instructors	Dr. Selehu Anteneh (salehu_anteneh@yahoo.com) [local coordinator, lead] Dr. Paul Gray () Dr. Solomon Negash (snegash@kennesaw.e.du, Skype: snegash) Dr. Ying Xie ()

1. Course Overview

This course will change the way you think about data and its role in business.

Businesses, governments, and society leave behind massive trails of data as a by-product of their activity. Increasingly, decision-makers rely on intelligent systems to analyze these data systematically and assist them in their decision-making. In many cases automating the decision-making process is necessary because of the speed with which new data are generated. This course connects real-world data to decision-making. Real-world examples from Finance, Marketing, and Operations are used to illustrate applications of a number of data mining methods. The use of real-world examples and cases places these techniques in context and teaches students how to avoid the common pitfalls of data mining, emphasizing that proper application of data mining techniques is as much an art as it a science. In addition to cases, the course features hands-on exercises with data mining software. The course is suitable for those interested in working with and getting the most out of data as well as those interested in understanding data mining from a strategic business perspective. It will change the way you think about data in organizations.

The goal of this course is three-fold. After taking this course you should:

1. ***Approach business problems data-analytically***. Think carefully & systematically about whether & how data can improve business performance.

2. ***Be able to interact competently on the topic of data mining for business intelligence.*** Know the basics of data mining processes, algorithms, & systems well enough to interact with CTOs, expert data miners, and business analysts. Be able to envision data-mining opportunities.

3. ***Have had hands-on experience mining data and applied research.*** Be prepared to follow up on ideas or opportunities that present themselves, e.g., by performing pilot studies

2. Instruction Method

This is primarily a lecture-based course, but student participation is an essential part of the learning process in the form of active technical/paper presentations and project discussion. The course will explain with detailed real-world examples the inner workings and uses of various data mining techniques. The first emphasis is on understanding the various types of data mining/business intelligence techniques, and when and how to use them, and secondarily on the mechanics of how they work. The final emphasis is on what it takes to produce high quality applied machine learning research paper.

Assignment Questions

Each class session has materials you must read prior to class. There will be a total of five question assignments, each comprising a (multi-part) question. The assignments will be posted in advance of the course. The assignments may involve hands-on work that will be completed in Weka—based on a data set that we will provide. You must turn in *all* question answers on the dates they are due. They will be graded and returned promptly. The goal of the assignments is to get you familiar with Weka-so that you may use Weka for your research papers. You should be able to use the computer labs for the assignments and ask questions of both the instructors and your colleagues to complete the assignments. You may also simply use your personal computers.

Late assignments

Turn in your assignment early if there is any uncertainty about your ability to turn it in on the due date. Assignments up to 1 day late will have their grade reduced by 50%. After one day, late assignments will receive no credit (no exceptions).

Research Paper

You are required to complete a novel research report (~15 pages)

You are required to format the document as if to submit the report for publication to a well known decision support systems or data mining journal. However, your grade will not depend on the success of your submission, only on its quality as assessed by the instructor. Your report should include Introduction, Related Work, Methodology, Results, Comparison / Analysis, Conclusion, Discussion and Future Work, and References. Your research report can be a *description of a new data mining method or algorithm you have developed*, or it can be an *analysis of a data set* toward answering an overarching applied research question.

Your instructors will help you with some project ideas and provide you with examples from prior years, though you are encouraged to choose your own, which will need approval. We will work on identifying novel research questions in the week prior to the course starting.

3. Requirements and Grading

This is a lecture-style course, however student participation is important. Students are required to be prepared and read the material before class. Students are required to attend all sessions and discuss with the instructor any absence from class. As discussed above, you will hand-in 5 (individual) write-ups to questions that will be assigned in class and will be posted on the class webCafe site. Answers should be well thought out and concise. Points will be deducted for sloppy language and irrelevant discussion.

There will be one team project in which students will address a real-world business problem with data mining/machine learning techniques. Students will hand in a report (accounts for 80% of project grade) and prepare a short class

presentation of their work (20% of project grade). A class discussion will follow the presentations. Details of the requirements for the project will be discussed the second day of class.

There will not be a final exam at the end of the semester.

The grade breakdown is as follows:

1. Assignment Questions (6 Write-ups): 30 points overall
3. Data Mining Project: 70 points (many parts to this)
4. Participation and Class Contribution: 10 points

4. Teaching Materials

The following are reading materials for this course:

1. Two Crows Corporation. Introduction to Data Mining: Third Edition.
   Available at: http://www.twocrows.com/intro-dm.pdf

2. Supplemental readings on the course website will be provided as the class progresses. Please let the instructor know immediately if any of the links are broken.

3. Website (pbwiki) for this course containing lecture materials and late breaking news, accessible through the course wiki page http://shawndra.pbwiki.com)

5. Course Schedule (Tentative Dates and Topics)

Segment	Topic	Due(Date)	Readings
Pre: Course Preparation	Personal Profile and Weka Installation	June 19 Hands On Assignment 1 Data Mining Profile	
Pre: Online Research June 15 – June 30	Literature Review and Research Question Selection	June 30 Written Assignment 1: 2 page proposal	
Face to Face: July 6 – July 17			
1	KDD – The Lay of The Land	July 6	Required Reading: Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth. "The KDD process for extracting useful knowledge from volumes of data", Communications of the ACM. Volume 39 , Issue 11 (November 1996). ACM Press New York, NY, USA Recommended Reading:

			R.O. Duda, P.E. Hart and D.G. Stork. "Pattern Classification", Wiley-Interscience, 2001. Chapter 1: Introduction
2	Some Methods for Classification Part I WEKA Lab	July 7 HW Due: Paper Summaries Hands on Assignment 2	Recommended Readings: R.O. Duda, P.E. Hart and D.G. Stork. "Pattern Classification", Wiley-Interscience, 2001. Chapter 8.1-8.4: Decision Trees Chapter 6: Neural Networks Tom Mitchell, Machine Learning, McGraw Hill, 1997. Chapter 6: Bayesian Learning Chapter 10: Learning Sets of Rules
3	Some Methods for Classification Part II	July 8 HW Due: Paper Summaries	Required Readings: Pedro Domingos and Michael Pazzani., On the Optimality of the Simple Bayesian Classifier under Zero-One Loss Machine Learning, 29, 103-130, 1997. Pdf Perlich, C., F. Provost, and J. Simonoff. "Tree Induction vs. Logistic Regression: A Learning-curve Analysis." To appear in the Journal of Machine Learning Research. CeDER Working Paper #IS-01-02, Stern School of Business, New York University, NY, NY 10012. Fall 2001. Pdf On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. Andrew Y. Ng and Michael Jordan. To appear in NIPS 14, 2002. ps, pdf Eibe Frank and Ian H. Witten (1998). Generating Accurate Rule Sets Without Global Optimization. In Shavlik, J., ed., Machine Learning: Proceedings of the Fifteenth International Conference, Madison, Wisconsin. Morgan Kaufmann Publishers, San Francisco, CA, pp. 144-151. Pdf
4	Genetic Algorithms/Association Rules	July 9 HW Due: Paper Summaries	Required Readings: Cooper, L. G., and G. Giuffrida Turning data mining into a management science tool: New

		Hands on Assignment 3	algorithms and empirical results, Journal of Management Science, 2000 (ps) Padmanabhan, B. and Tuzhilin, A.,Small is Beautiful: Discovering the Minimal Set of Unexpected Patterns, (pdf)) 2000. Procs. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining pages 54-64, August 2000. Padmanabhan, B., Zheng, Z., and Kimbrough, S., Personalization from Incomplete Data: What you don't know can hurt, PDF, 2001 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining Zheng, Z., Padmanabhan, B., and Kimbrough, S., On the Existence and Significance of Data Preprocessing Biases in Web Usage Mining, PDF, 2002.
5	Evaluation	July 10 HW Due: Paper Summaries	Required Readings: Provost, F. and T. Fawcett, "Robust Classification for Imprecise Environments." Machine Learning 42, 203-231, 2001. Pdf D. Jensen and P.R. Cohen. "Multiple comparisons in induction algorithms", Machine Learning 38(3) 1999. Pdf R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In IJCAI 1995, 1995.pdf.
6	OLAP/Traditional Tools	July 13 HW Due: Paper Summaries Hands on Assignment 4	TBD
7	Relational Learning	July 14 HW Due: Paper Summaries	Required Readings: R. Quinlan and R.M. Cameron-Jones. Induction of logic programs: FOIL and related systems. New Generation Computing, 13:287 312, 1995. I4. Sean Slattery and Mark Craven.

			Combining statistical and relational methods for learning in hypertext domains. In Proceedings of the 8th International Conference on Inductive Logic Programming, ILP-98, pages 38 52, 1998. Neville, J., D. Jensen, L. Friedland and M. Hay. (2002) Learning Relational Probability Trees. University of Massachusetts Amherst, Technical Report 02-55. Revised version February 2003. D. Koller and A. Pfeffer. Probabilistic frame-based systems. In Proc. AAAI, 1998. http://citeseer.nj.nec.com/koller98probabilistic.html Perlich, C. and F. Provost. "Aggregation-Based Feature Invention for Relational Learning." Under Review for SIGKDD 2003. Preliminary version: CeDER Working Paper #IS-03-03, Stern School of Business, New York University, NY, NY 10012. Spring 2003.
8	Text Mining/Business Intelligence	July 15 HW Due: Paper Summaries	Recommended Reading: R. Baeza–Yates and B. Ribeiro-Neto. Modern Information Retrieval, Addison-Wesley 1999. Chapter 2.5:Classic Information Retrieval Chapter. 7; Text Operations Required Readings: Daniel Billsus and Michael Pazzani, "A Personal News Agent that Talks, Learns and Explains", Proceedings of the Third International Conference on Autonomous Agents, 1999. http://citeseer.nj.nec.com/billsus99personal.html Dwi H. Widyantoro and Thomas R. Ioerger and John Yen, "An Adaptive Algorithm for Learning Changes in User Interests", Proceedings of the Eigth International Conference on Information and Knowledge Management, 1999. http://citeseer.nj.nec.com/523103.html

			Good, N., Schafer, J.B., Konstan, J., Borchers, A., Sarwar, B., Herlocker, J., and Riedl, J., Combining Collaborative Filtering with Personal Agents for Better Recommendations. Proceedings of the 1999 Conference of the American Association of Artificial Intelligence (AAAI-99). http://citeseer.nj.nec.com/good99combining.html
9	Technical IS Research Methods/Applied Data Mining in Developing Nations	July 16 HW Due: Paper Summaries Hands on Assignment 5	Required Readings: Hevner, A., S. March, J. Park, and S. Ram, "Design Science Research in Information Systems," Working Paper, Carlson School School of Management, University of Minnesota, Minneapolis, MN, 2001. (most recent version available from Al Hevner).pdf Weber, R., "Toward a Theory of Artifacts: A Paradigmatic Base For Information Systems Research," Journal of Information Systems, Spring, 1987.PDF Langley, P., "Crafting Papers on Machine Learning," Journal of Information Systems, Spring, 1987. HTML
10	Wrap Up - Draft Paper Presentations	July 17 Constructive criticism on project presentations	
Post: Assignment Submissions		July 31 Literature Review: Summarizing current literature on Data Mining/Machine Learning in developing countries August 14 Hands on Assignment 6: Comparison of methods on given dataset	
Post: Online Practice	Term Paper writeup and workshop Term Paper submission	August 28 Final Term Paper	

Summary of Tentative Course Schedule

Segment	Date	Topic	Due
Online Research	June 15-30	Literature Review and Research	2 page proposal
Course Preparation	June 20	Personal profile and WEKA installation	Hands on Assignment 1
Face-to-Face	July 6- July 17		
	July 6	KDD – The Lay of the Land	
	July 7	Some Methods for Classification Part I	HW: Paper Summaries Hands on Assignment 2
	July 8	Some Methods for Classification Part II	HW: Paper Summaries
	July 9	Genetic Algorithms / Association Rules	HW: Paper Summaries Hands on Assignment 3
	July 10	Evaluation/ WEKA Lab	HW: Paper Summaries
	July 13	OLAP/Traditional Tools/Project Update	HW: Paper Summaries Hands on Assignment 4
	July 14	Relational Learning	HW: Paper Summaries
	July 15	Text Mining/Business Intelligence Privacy Preserving Data Mining	HW: Paper Summaries
	July 16	Technical IS Research Methods/ Applied Data Mining in Developing Nations	HW: Papers Summaries Hands on Assignment 5
	July 17	Wrap Up/ Draft Paper Presentations	Peer Review
Post: Assignments	July 31		Hands on Assignment 6
	Aug 14		Literature Review
Post: Online Practice	Aug 31	Term Paper Writeup and workshop	Term Paper

Expected student-professor interaction during Online Research

E-mail Q&A	Q&A in topic selection
Resource inquiry	Request for additional journal articles
Review and Feedback	Preliminary review and feedback on student paper

Expected student-professor interaction during face-to-face

Class Session	30 hours	Lecture and discussion
Simulation	10 hours	Hands-on WEKA
One-on-one	15 hours	Mentoring and feedback on term paper

Expected student-professor interaction during Online Practice

Forum Discussion	Student-instructor and student-student interaction via forum
Student Workshop	Constructive critic among students; presentation at student-led workshop

Paper Submission Final term paper submission