APPLIED DATAMINING ONLINE CLASS

INSTRUCTOR: HASAN KURBAN

August 21, 2017

CONTENTS

LIST OF TABLES

* *Department of Computer Science, School of Informatics and Computing, Indiana University, Bloomingon, IN, USA.*

1 SYLLABUS

Datamining is the *process* of discovering novel, latent, useful, and actionable information from usually large corpuses of data. Conducted virtually everywhere, datamining is still as much an art (experience) as straight computation. Datamining is a fascinating endeavor that potentially has enormous payoffs. This *living* document will be dynamically updated to reflect new material, errata, and so forth.

This is a graduate class in the Data Science Program of the School of Informatics and Computing, Indiana University, Bloomington, IN, USA. The learning objective is to broadly familiarize students with the elements of data mining. Students are expected to be profficient in algebra, to have familiarity with probability and calculus. While the student is expected to be proficient in R, the first week will a refresher. Additionally, since there will be a fair amount of writing, LaTeX will be required, a type-setting language, that produces professional documents used for books, ebooks, *etc.* Although there are many freely available, MikTeX is strongly suggested for its ease of use.

The class requires one textbook, R, and and a LaTeX compiler:

- Text "Data Mining with R: Learning with Case Studies," 2^{nd}., by L. Torgo, Chapman and Hall.
- R https://www.r-project.org/
- LaTeX https://miktex.org/

LaTeX is easy and fun to learn and will become an invaluable tool. This document itself is written in LaTeX. Lastly, both seminal and new papers as well as current and timely new stories may occassionally be added for reading.

The remainder of the syllabus contains contact information, grading policy, and an outline of the course.

1.1 Contact Information

zoom	https://zoom.us/j/6103564064
office hours	Wednesday-Friday 10:00A-11:00A EST or by appointment.
email	hakurban@indiana.edu

1.2 Learning Outcomes

There are four major learning outcomes. They are co-equal in importance.

Knowledge Area
Overall Datamining Process
Elements of the Process
Machine Learning Algorithms
Interpretation of Datamining

More plainly, the student should be able to assess a potential datamining problem, employ the process, which includes the appropriate algorithm, interpret the results, and suggest an outcome clearly and succintly.

1.3 Calendar

IUB's official calendar is found here:
http://registrar.indiana.edu/official-calendar/official-calendar-fall.shtml
Important dates are:

Begins	Mon	Aug 21
Labor Day	Mon	Sept 4
Fall Break		Oct 6-8
Auto W	Sun	Oct 22
Thanksgiving		Nove 19-26
Final Exams		Dec 11-15
Ends	Fri	Dec 15

1.4 Awarding of Grades

There will be two exams, 10 homeworks, and a cumulative final. If a student fails the final, then the student will not pass the class. Grades are based on straight scale 99-100 A+, 93-98 A, 90-92 A-, 88-89 B+, 83-87 B, 80-82 B-, 78-79 C+, *etc.* Grades *may* be positively curved upon the discretion of the instructor. The grade breakdown is given below:

Element	%	Points
Homework	50	200 pts. each
Exams	20	100 pts. each
Final	30	200 pts. (cumulative)

Formally, the semester grade $g(h, e_1, e_2, f)$ is:

$$g(h, e_1, e_2, f) = .5h + .2(e_1 + e_2) + .3f \quad (1)$$

where h, e_i, f are the ratio of points received to maximal points that can be earned. For example, if the exam is maximally worth 100 points and 80 points are earned, then $e = \frac{80}{100} = .8$.

- *Cheating* of any kind will not be tolerated. Students are invited to visit https://studentaffairs.indiana.edu/office-student-ethics/misconduct-charges/academic-misconduct.shtml. Quoting the relavent passage:

  > ... Coursework performed while misconduct proceedings are underway, however, shall be considered conditional. Conditional work may be affected or eliminated based on a final finding of misconduct or sanction imposed. This may result in loss of course credit, a delay in the awarding of a degree, or revocation of a degree that was awarded prior to a final decision in the misconduct proceedings. If either academic or personal misconduct is discovered that may impact degree conferral or graduation, the Dean of Students may notify the student's academic dean, who may withhold conferral of the degree pending completion of misconduct proceedings.
  >
  > If, after a degree has been conferred, the University determines that the student committed academic misconduct prior to the conferral, the University may revoke the degree ...

- Homework must be thoughtfully executed: clear, concise, free of grammatical and spelling errors.
- Homework will explicitly state **INPUTS**: the readings, extra materials to read, files to be used and **OUTPUT**: Deliverables–what to turn-in. All homework will contain at least a PDF and the Tex file used to create it. Additionally, students may be required to submit both the logs and R-scripts. Reiterating, all products must be the student's alone. Explictly state libraries used.
- If, "All the work herein is solely mine," is not explicitly stated as a preamble to the homework, then it will not be graded.
- A sample exam one, exam two, and final annotated with some solutions will be made available before the scheduled exams.

1.5 Tentative Schedule

The following is a tentative outline of the materials. There are 10 homework–most chapters having two. Students will be provided the LaTeX source code to more easily turn-in homework. The first outline is the process of the class and the second is a more detailed outline of content from the book.

- Week 1
  - Introduction to the Class
  - Lecture 1
  - Chapter 2:7-41
  - Homework 1
- Week 2-3
  - Datamining as a Process and its Parts: Data Acquisition, Cleaning, Enrichment, Transformation, Machine Learning, Interpretation.

    This section introduces the student to each of these processes.
  - Lecture 2, 3
  - Chapter 3:43-165
  - Homework 2.1
  - Homework 2.2
- Week 4-6
  - Datamining Algal Data

    > Algae–microscopic plantkon–is a critical food source for many kinds of shellfish. Their over abundance (*blooms*) can have, however, devistating affects on oceanic ecosystems through production of toxins and overuse of oxygen[a]. In this case, we use datamining to predict these blooms. The preprocessing of data in datamining, it is widely held, requires 80% of the resources of a project. In this case, data must be heavily preprocessed.
    >
    > *a* http://www.noaa.gov/what-is-harmful-algal-bloom
  - Lecture 4-6
  - Chapter 4:193-249
  - Homework 4.1
  - Homework 4.2
- Exam 1
  - All material from Lecture, Homework, and Text
  - Lecture 7 (review)
  - Students will be asked to write R code as well
  - Exam is in-class 1.5 hours closed book
- Week 5-7

  - Business: Risk

    In Finance, risk is bought or sold, and success is simply financial gain as a function of risk. Stocks are a form of risk, and success is to sell at a higher price than their purchase price. The value of stock, however, is not easily determined beyond a very small window. In this case, the student will build a model to improve decision making on the buying and selling of stock[a]. Several popular approaches are used in this case, including neural nets (NN), suport vector machines (SVM), and splines. Patently, this is not an endorsement to invest–it simply yields a different domain for students to use datamining.

    a http://www.nasdaq.com/investing/start-investing-1000.stm

  - Lecture 5-7
  - Chapter 5:241-292
  - Homework 5.1
  - Homework 5.2

- Week 8-10
  - Security: Detecting Fraud (Outlier Analysis)

    Currently the moniker “fake news" has become widely applied to information that is not valid. More broadly, fraud is attempting to make what is false appear true, usually for financial gain[a]. The challenge becomes determining what does *not* fit a pattern. This case study uses several popular techniques, *e.g.*, Naïve Bayes and AdaBoost to determine fraud. Additionally, important valuation measures are presented–like Life Charts.

    a https://www.fbi.gov/scams-and-safety/common-fraud-schemes/business-fraud

  - lecture 8-10
  - Chapter 6:295-350
  - Homework 6.1
  - Homework 6.2

- Exam 2
  - All material from Lecture, Homework, and Text
  - Lecture 11 (review)
  - Students will be asked to write R code as well
  - Exam is in-class 1.5 hours closed book

- Week 10-12
  - Genes Expession Levels

> Genes–stretches of DNA–are the blueprints for proteins that, in turn, guide biological function. Microarrays measure the intermediary, RNA, whose proportion will be proportional to the proteins each encodes. This set of expressions can be used to classify people, say, with an illness [a]. One of the most powerful classifiers is an ensemble–a collection of weak classifiers. In this case study, Random Forests are used which perform quite well in a wide range of domains.
>
> a https://www.nature.com/scitable/topicpage/gene-expression-14121669

  - lecture 11-12
  - Chapter 7:353-381
  - Homework 7.1

- Final (Cumulative)
  - All material from Lecture, Homework, and Text
  - Lecture 13 (review)
  - Students will be asked to write R code as well
  - Final is 3 hours in-class closed book

1.5.1 *Detailed Outline of the Topics, in order, of the class*

Since students are familiar with R, we are skipping Chapter 1 and reviewing quickly some of the most commonly used elements in R.

I R and Data Mining

Chapter 2 Introduction to R

2.1 Starting with R
2.2 Basic Interaction with the R Console
2.3 R Objects and Variables
2.4 R Functions
2.5 Vectors
2.6 Vectorization
2.7 Factors
2.8 Generating Sequences
2.9 Sub-setting
2.10 Matrices and Array
2.11 Lists
2.12 Data Frames
2.13 Useful Extensions to Data Frames
2.14 Objects, Classes and Methods
2.15 Managing Your Sessions

LaTeX Your first LaTeX Document

Chapter 3 Introduction to Data Mining

3.1 A Bird's Eye View on Data Mining
3.2 Data Collection and Business Understanding