EE 361M Introduction to Data Mining (and Machine Learning and Data Science)

Course description

The design and application of predictive models is a central goal of datamining and machine learning. In this course, we will study a variety of techniques for predictive modeling. Particular emphasis will be given to approaches that are scalable to very large datasets and/or those that are relatively robust when faced with a large number of predictors, and algorithms for heterogenous or streaming data. Many of these capabilities are essential for handling BIG DATA. Connections to relevant business problems shall be made via example studies. We will mostly be using Python and R for statistical modeling.

The main goal of this course is to convey and understanding of the pros and cons of different predictive modeling techniques, so that you can (i) make an informed decision on what approaches to consider when faced with real-life problems requiring predictive modeling, (ii) apply models properly on real datasets so as to make valid conclusions. The goal will be reinforced through both theory and hands-on experience.

The tentative schedule of classes and covered topics can be found here.

Term project guidelines will be posted here (/ghosh/teaching/361m-sp16/projects).

Information about the course instructor and TA(s) is available on the contact page (/ghosh/teaching/361m-sp16/contact).

PREREQUISITES:

This course is meant for junior/senior level ECE and CS students. You MUST have taken (with grade of C- or better)
EE351K (or equivalent): Probability and Statistics
AND M340L (or equivalent): Matrices/Linear Algebra
AND at least one data structures/algorithms course such as EE422C, EE360C or equivalent.

For any exceptions to the above pre-reqs, you need written consent from me (please send me a resume and a para describing why you have adequate background for this class), else you will be dropped from the class. Graduate students are not allowed in this class. (Integrated BS/MS students from ECE/CS are OK).

Textbooks

The material for the lectures is taken from a wide variety of sources. My notes will be available via Canvas. The Scikit-Learn (http://scikit-learn.org/stable/) website serves as the key “textbook”. Two other, more traditional textbooks for the course are:

Author: Max Kuhn and Kjell Johnson (**KJ**)
Title: Applied Predictive Modeling
Publisher: Springer
ISBN: 1461468485
Year: 2013
Notes: Available through Amazon, Springer and UT Co-op.

Author: Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (**JW**)
Title: An Introduction to Statistical Learning with Applications in R
Publisher: Springer
Notes: The authors have kindly provided a free pdf version here (http://www-bcf.usc.edu/~gareth/ISL/ISLR%20First%20Printing.pdf), though it may be worth your while to get a hardcopy as well.

Relevant sections of these books are indicated next to the topics below. A more detailing listing of readings/videos/other resources is provided through Canvas.
Topics without reference to KJ, JW will be covered through lecture notes and a reading list of papers (also see supplementary texts at the bottom).

Course Schedule and Topics

KJ, JW refers to the two texts. HTF and B refer to two of the (more advanced) supplementary books and they are serve as references for the interested student.

1. Overview: The data mining process; model fitting and overfitting; decision theory; probability review; Types of predictive analytics; Local vs. global models; Tackling Big Data; Multiple linear regression.

(3 lectures; KJ: Chapter 1,2; JW Chapters 1-3, B, Ch 1, 2.1-2.3; HTF Ch 1, 2.1-2.6)
Objective: Provide overview and context for this class.

2. Data Pre-Processing: Cleaning, Reduction, Feature Extraction and Visualization: Data quality; Curse of dimensionality, Transformations, Imputation, Sampling, Outlier detection, PCA.

(3 lectures, KJ: Chapter 3, JW 10.1, 10.2, B 12.1; HTF Ch 14.5, 14.8)
Objective: Understand that good data quality is a pre-requisite for effective models, and study some methods for improving data quality.

3. Regression: generalized regression, bias-variance tradeoff and overfitting; Model tuning; Basis function expansion; Dealing with large number of features; Ridge, Lasso and Stagewise Approaches; Non-linear methods;

(3 lectures; KJ: Chapter 4-6, 7.1,7.4, 7.5; JW Ch 6, B 3.1, 3.2; HTF Ch 2.7, 2.8, 3.1-3.4, 7.1-7.3, 11.1-11.8)

Objective: Learn to understand predictive models where desired outcome is a numeric quantity.

4. Classification: Bayes decision theory, Naïve Bayes and Bayesian networks; (Scaled) Logistic regression; LDA; Scaling decision trees to big data; Kernel methods and Support Vector Machines (SVMs) for classification and regression; dealing with class imbalance

(6 lectures, KJ: Chapter 11,12,13,14.1,14.2,8.1,16; JW 4, 8, 9; B 4.1-4.3.4; 6.1, 6.2, 7.1, 14.4; HTF Ch 4, 7.10, 9.2, 12, 13.3)

Objective: Learn to build and evaluate predictive models where desired outcome is a class label.

5. Clustering and Co-clustering: k-means; hierarchical methods, graph partitioning; co-clustering, semi-supervised learning. Market Basket applications

(3 lectures: JW 10.3, 10.5, B 9.1, 9.2; HTF Ch 13.1, 13.2, 14.3, 14.4)

Objective: Learn issues involved in unsupervised learning and the trade-offs among alternative approaches to clustering.

6. Ensemble Methods: Model Averaging, Bagging and Random forests, boosting, Gradient boosting; Bag of Little Bootstraps

(2 lectures; KJ: Chapter 14.3-14.8; JW 8.2; B 14.2, 14.3, HTF Ch 8.7, 8.8, 10.1-10.7, 16)

Objective: Understand the benefits of combining multiple predictive models.

7. Streaming Data Mining: Online learning – basic approaches; Winnow/Voted Perceptrons; Stochastic gradient methods for large data sets

(1 lecture)

8. Semi-supervised Learning for Big Data: Learning when labeled data is scarce.

(1 lecture)

Objectives for 7 and 8: Both streaming and semi-supervised situations are commonly involved in big data applications, and so ways of predictive modeling in these non-traditional settings are covered

9. Specialized Topics Topics: (coverage depends on time available and interest of class)

Deep learning, Recommender Systems; Customer-product affinity detection;Hadoop/SPARK, Azure ML

10. Term Project Presentations and Discussion

(4 classes)

11. Wildcards: A couple of classes may be used for invited talks by visiting experts.

Grading information

- 10+25%: Project (/ghosh/teaching/361m-sp16/projects) (groups of 3-5): (project outline + 2 presentations) + term paper due May 11th
- 30%: 5 Asssignments
- 15%: 3 pop-quizzes
- 20%: Written Exam (Tues March 29, in class)

There will be no final exam.

Quizzes will be held in class and of duration 15-20 minutes. Their objective is to review key concepts introduced in class.

At the end of the course, you will get a score out of 100 based on the percentages stated above. Your final grade will be solely based on this score. The grade is primarily based on the curve, i.e. is relative to how the whole class performs; however entire curve may shift up or down a bit depending on how the class as a whole performs relative to past classes. **Grading is NOT based on absolute thresholds, e.g. 90+ = A etc.**

Supplementary Texts

Author: Wes McKinney
Title: Python for Data Analysis
Publisher: O'Reilly

MOOC: Coursera course by Andrew Ng has some very introductory material on linear algebra, e.g. multiplying a matrix with a vector.
Notes: https://class.coursera.org/ml-003/lecture (https://class.coursera.org/ml-003/lecture)

Author: Trevor Hastie, Robert Tibshirani, and Jerome Friedman (HTF)
Title: The Elements of Statistical Learning
Publisher: Springer (2nd edition)
ISBN: 0387848576
Notes: Can get it from Amazon, about $70 but well worth it, or download pdf from http://www-stat.stanford.edu/~tibs/ElemStatLearn/ (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)

Author: Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (TSK)
Title: Introduction to Data Mining
Publisher: Addison-Wesley (2005)
ISBN: 0-321-32136-7
Notes: Some chapters are downloadable from this website (http://www-users.cs.umn.edu/~kumar/dmbook/index.php)

Author: Christopher M. Bishop (B)
Title: Pattern Recognition and Machine Learning
Publisher: Springer
ISBN: 0387310738
Notes: http://research.microsoft.com/en-us/um/people/cmbishop/prml/ (http://research.microsoft.com/en-us/um/people/cmbishop/prml/)

Author: Kevin Murphy
Title: Machine Learning: A Probabilistic Perspective,
Publisher: MIT Press
ISBN: 0262018020
Notes: Covers a very wide range of topics. Lots of examples in Matlab, with source code access.

Disabilities statement: "The University of Texas at Austin provides upon request appropriate academic accommodations for qualified students with disabilities. For more information, contact the Office of the Dean of Students at 471-6259, 471-4641 TTY."

NOTICES:

- Students with disabilities may request appropriate academic accommodations from the Division of Diversity and Community Engagement, Services for Students with Disabilities, 471-6259, http://www.utexas.edu/diversity/ddce/ssd/ (http://www.utexas.edu/diversity/ddce/ssd/)
- A notice regarding academic dishonesty. UT Honor Code and example of what constitutes plagiarism : http://registrar.utexas.edu/catalogs/gi09-10/ch01/index.html (http://registrar.utexas.edu/catalogs/gi09-10/ch01/index.html)
- A notice regarding accommodations for religious holidays. “By UT Austin policy, you must notify me of your pending absence at least fourteen days prior to the date of observance of a religious holy day. If you must miss a class, an examination, a work assignment, or a project in order to observe a religious holy day, you will be given an opportunity to complete the missed work within a reasonable time after the absence.”)
