Principles of Data Mining

Dr. Anil Chaturvedi
anilchaturvedi@uchicago.edu

This course is aimed at providing students a rigorous methodological foundation in analytical and software tools to successfully undertake projects involving Data Mining. Students are exposed to concepts of exploratory analyses for uncovering and detecting patterns in multivariate data, hypothesizing and detecting relationships among variables, conducting confirmatory analyses, and building models for predictive purposes. It will present predictive modeling in the context of balancing predictive and descriptive accuracies.

Students will learn applications of statistical sampling and inference, conducting exploratory analyses via single-mode and multi-mode cluster analyses, performing dimension reduction via two-mode (principal components and factor analysis models) and higher-mode (Three-mode) hybrid factor analysis , analysis of longitudinal correlations via three-way models, predictive modeling via Logistic Regression, Multinomial Logit, and Classification and Regression Trees. Students will also be exposed to concepts of cluster-wise regressions and latent-class analysis. Students will learn about model selection criteria – in both training samples (criteria such as Likelihood based Chi-Squares, AIC, BIC), and holdout samples (criteria such as Holdout accuracies).

At the end of the course, students would be able to independently perform both exploratory and confirmatory analyses, and conduct data mining for exploratory and predictive purposes.

Textbooks

[ISLR] An Introduction to Statistical Learning, with Applications in R, Gerath James, Daniella Witten, Trevor Hastie and Robert Tibshirani, Springer, ISBN 978-1-4614-7137-0 ISBN 978-1-4614-7138-7 (eBook). Available for free download from Internet.

Quantitative Models in Marketing Research, Philip Hans Franses and Richard Paap, Cambridge University Pres,, 2007, Pages 76-86 only, Chapter 5. ISBN: 978-0-521-80166-9. Library Course Reserves

Data Mining and Analysis: Fundamental Concepts and Algorithms, Zaki, M. J. and Wagner Meira, Chapter 8, Pages 217-237. ISBN 978-0-521-76633-3, Cambridge Press. Library Course Reserves

Grading

Homework Assignments	50% (1+4 Biweekly assignments – Due Sessions 3, 5, 7, and 9)
Paper Review and Presentation	10%
Class Presence, Participation, & Engagement	10%
Project	30%
Bonus Assignments (2)	10%

Schedule

Session 1	Course Setup. Project Selections. Introduction to Business Analytics, Statistical Learning and Data Mining: Exploratory vs Confirmatory analyses. Multivariate models
Session 2	Data Reduction and Pattern detection via partitioning: Big Data Algorithms
Session 3	Data Reduction and Pattern detection: Clustering Categorical Data: Latent Class and K-Modes
Session 4	Feature Extraction and Dimension Reduction: Principal Components and Factor Analysis

Session 5	Predictive Modeling via Logistic Regression and Multinomial Logit
Session 6	Predictive Modeling via Classification and Regression Trees; CHAID
Session 7	Cluster-wise Regressions
Session 8	Discriminant Analysis
Session 9	Data Mining using Association Rules
Session 10	Project Presentations and Reports

Assignments

1. Students work with actual Data sets suggested to them to perform analyses
2. Statistical Language for analysis: R

Session 1: Course Set up; Syllabus and Grading. Introduction to Statistical Learning and Data Mining. Exploratory and Confirmatory analyses. Multivariate Models.

Required Reading

1. Chapter 2, Pages 15-51. ISLR

Recommended Reading

1. Carroll, J. D., and Chaturvedi, A. (1995). A General Approach to Clustering and Multidimensional Scaling of Two-way, Three-way, or Higher-Way Data. In R. D. Luce, M. D'Zmura, D. D. Hoffman, G. Iverson & A. K. Romney (Eds.), *Geometric Representations of Perceptual Phenomena* (pp. 295–318). Mahwah, NJ: Erlb (T) **(Methods) Library Course Reserves [On Chalk – Title CANDCLUS]**
2. **Data Mining: What Is It?**

Session 2: K-means Clustering: Big Data Algorithms: Overlapping Clustering

Required Reading:

1. Chapter 10, Pages 385-399 and Pages 404-413. ISLR
2. Chaturvedi, Anil D. Big Data Marketing Analysis. Challenges and Proposed Solutions. DMA Analytics Journal, Pages 37-45. [On Chalk]

Recommended Reading:

1. Hedge Fund Classification Using K-Means Clustering Method, Nandita Das, 2011. **(Application)** Link to Paper
2. M.Suresh Babu, Dr. N.Geethanjali, Prof B.Satyanarayana Clustering Approach to Stock Market Predictions, Int. J. Advanced Networking and Applications, Volume: 03, Issue: 04, Pages:1281-1291 (2012). **(Application)**
3. MacQueen, J. B. (1967), "Some Methods for Classification and Analysis of Multivariate Observations," in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, L. M. LeCam and J. Neyman, eds. 281-97. aum. **(Methods)**
4. Vinod, H. D. (1969), "Integer Programming and Theory of Grouping," Journal of the American Statistical Association, 64, 506-19. **(Methods)**

5. Chaturvedi, A.D, Carroll, J. D., Green, P. E., and Rotondo, J. A. (1997). A Feature-Based Approach to Market Segmentation via Overlapping K-Centroids Clustering. *Journal of Marketing Research*, XXXIV, 370–377. **(Methods)**
6. Parmar, P.S. and Pandi(Jain) G.S. (2015), Performance Analyis and Augmentation of K-Means Clustering, based Approach for Human Detection in Videos, IJEDR, Volume 3, Issue 2, 1029-1035. **(Methods and Application)**
7. Chen Y. and Tu, L. (200X), Density-Based Clustering for Real-Time Stream Data, ACM **Copyright (Methods and Application)**. Reference incomplete.

Session 3: Clustering Categorical Data: K-Modes and Latent Classes

Required Reading:

1. Chapter 10, Pages 385-399 and Pages 404-413. ISLR
2. Illustration of Local Independence Assumption

Recommended Reading:

1. Chaturvedi, A. D., Green, P. E., and Carroll, J. D. (2001). K-Modes Clustering. *Journal of Classification*, 18, 35-56. **(Methods)**
2. Lanza, S.T., Savage, J.S., and Birch, L.L. (2012), Identification and Prediction of Weight Loss Strategies Among Women, Obesity A Research Journal, Volume 18, Issue 4, 833-840. **[Methods and Application]**
3. Rindskopf, D. (20**), The Use of Latent Class Analysis in Medical Diagnosis, Joint Statistical Meetings-Social Statistics Section.2912-2916.[**Application]**
4. poLCA: An R Package for Polytomous Variable Latent Class Analysis, Drew A. Linzer and Jeffrey B. Lewis, Journal of Statistical Software, Volume VV, Issue II, Pages 1-28. READ PAGE 3-7 FOR THIS MODULE**(Methods)**

Session 4: Principal Components and Factor Analysis

Required Reading:

1. Chapter 10, Pages 373-385 and Pages 401-404. ISLR

Recommended Reading:

1. Forecasting with Principal Components Analysis: An Application to Financial Stability Analysis, Mingione **(Application)**
2. http://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/ **(Application)**

Session 5: Logistic Regression and Multinomial Logit

Required Reading:

1. Chapter 4, Pages 127-137 and Pages 154-161. ISLR

Recommended Reading:

1. Kim A Keating and Steve Cherry, **Use and Interpretation of Logistic Regression in Habitat Selection Studies**, Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA **(Application)**
2. Wine-Tasting Examples, Wine-Tasting Study **(Application)**

3. Applied Logistic Regression, Third Edition, Hosmer, Lemeshow, and Sturdivant, Chapters 1-3 (Chapter 4 Optional). ISBN: 978-0-470-58247-3. **(Methods)**
4. Quantitative Models in Marketing Research, Philip Hans Franses and Richard Paap, Cambridge University Press, 2007, Pages 76-86, Chapter 5. ISBN: 978-0-521-80166-9. **(Methods)**
5. McFadden, D. (1974) "The Measurement of Urban Travel Demand", *Journal of Public Economics*, 3, pp. 303-328. **(Methods)**
6. McFadden, D. (1973) Conditional Logit Analysis of Qualitative Choice Behavior, in P. Zarembka ed., *Frontiers in Econometrics*, New-York: Academic Press. **(Methods)**

Session 6: Classification and Regression Trees

Required Reading:

1. Chapter 8, Pages 303-316 and Pages 323-328. ISLR

Recommended Reading

1. Iván Cantador, Desmond Elliott, Joemon M. Jose **A Case Study of Exploiting Decision Trees for an Industrial Recommender System**, Department of Computing Science, University of Glasgow, Lilybank Gardens, Glasgow, G12 8QQ, UK **(Application)**
2. Bruno Carneiro da Rocha1,2 and Rafael Timóteo de Sousa Júnior**, Identifying Bank Frauds using Crisp-DM and Decision Trees**, International journal of computer science & information Technology (IJCSIT) Vol.2, No.5, October 2010 **(Application)**
3. Classification and Regression Trees, Breiman, Friedman, Olshen, and Stone, Chapters 1-2, Chapters 8-9. ISBN: 0-412-04841-8. **(Methods)**
4. "An Introduction to Recursive Partitioning, Using the Rpart Routines", Terry M. Therneau and Elizabth J. Atkinson. **(Methods)**
5. Software Review: SPSS for Windows CHAID 6.0, Chicago: SPSS Inc, 1992, Anil Chaturvedi and Paul Green, 1992, Journal of Marketing Research, pp 245-254. **(Methods)**

Session 7: Cluster-wise Regression

Required Reading:

1. **Why Use Latent Class Regression? Radius Report.**

Recommended Reading

1. Astrid D. A., Kemperman M, and Harry J. P. Timmermans: Preferences, Beneftis, and Park Visits: A Latent Class Segmentation Analysis, *Tourism Analysis*, Vol. 11, pp 1-10. **(Application)**
2. Amit Bhatnagar and Sanjoy Ghose, A Latent Class Segmentation Analysis of E-Shoppers, Journal of Business Research, 57, 758-767, 2004. **(Application)**
3. Spath, H. (1979), "Algorithm 39: Clusterwise Linear Regression," *Computing, 22,* 367-373**(Methods)**
4. DeSarbo W. S. and Cron, W. L. (1988), A Maximum Likelihood Methodology for Clusterwise Regression, Journal of Classification, pp 249-282. **(Methods)**

Session 8: Discriminant Analysis

Required Reading:

1. Chapter 4, Pages 138-168. ISLR

Recommended Reading

1. Manuel Artis, Monsterrat Guillen, and Jose M. Martinez, A Model of Credit Scoring: An Application of Discriminant Analysis, Questiio, Volume 18, 3 pp, 385-395, 1994. **(Application)**
2. Mirko Savic, Dejan Brcanov, and Stojanka Dakic, Discriminant Analysis: Applications and Software Support, Management Information Systems, Volume 3, Number 1, 2008, pp 029-033. **(Application)**
3. Alka Brahmandkar, Discriminant Analysis: Applications in Finance, The Journal of Applied Business Research, Volume 5, Number 2, p 37-41. **(Application)**

Session 9: Association Rules

Required Reading:

1. Data Mining and Analysis: Fundamental Concepts and Algorithms, Zaki and Meria, Chapter 21, Pages 217-237. ISBN 978-0-521-76633-3. **(Methods)**

Recommended Reading

1. Tackett, James. Association Rules for Fraud Detection, Journal of Corporate Accounting and Finance, Volume 24, Issue 4, pages 15-22, June 2013. **(Application)**
2. Lucas Lau and Arun Tripathi, Mine your Business – A Novel Application of Association Rules for Insurance Claims Analytics, Casualty Actuarial Society E-forum,2011, Winter. **(Application)**
3. Todd Wittman, Time Series Clustering and Association Analysis of Financial Data, CS8980 Project. **(Application)**
4. http://www.statsoft.com/textbook/association-rules

Session 10: Final Project Presentations

Final Project Presentations and Project Reports Due

Projects

Students will complete a Data Analysis project. They will select their own data set – either from their work background, or from other publicly available data sources.

Late Work

All assignments must be submitted to the Chalk site for the course on the due date before 11:59 pm. If you turn in an assignment late, 10% credit will be deducted from the total score for each day after the deadline. Assignments turned in more than one week late will not receive credit. In the case of unexpected events, you must contact the instructor before the assignment due date in order to receive a grace period. Students can only receive up to two grace periods in the course.

Requesting Reasonable Accommodations

If you are interested in requesting disability accommodations, you may want to begin by reading through the information published on this website https://disabilities.uchicago.edu/. Also, please do communicate your requests as soon as possible to Gregory Moorehead, director of disability services, at 773.702.7776 or gmoorehead@uchicago.edu.

Academic Honesty and Plagiarism

It is contrary to justice, academic integrity, and to the spirit of intellectual inquiry to submit another's statements or ideas of work as one's own. To do so is plagiarism or cheating, offenses punishable under the University's disciplinary system. Because these offenses undercut the distinctive moral and intellectual character of the University, we take them very seriously.

Proper acknowledgment of another's ideas, whether by direct quotation or paraphrase, is expected. In particular, if any written or electronic source is consulted and material is used from that source, directly or indirectly, the source should be identified by author, title, and page number, or by website and date accessed. Any doubts about what constitutes "use" should be addressed to the instructor.

Data Sets: Students can get their own data. Hera are some useful links:

http://archive.ics.uci.edu/ml/ (RECOMMENDED)

http://www.mldata.org
http://www.dmoz.org/Computers/Artificial_Intelligence/Machine_Learning/Datasets/
Data Sets from rdatamining.com

http://homepages.inf.ed.ac.uk/rbf/IAPR/researchers/MLPAGES/mldat.htm
http://kdd.ics.uci.edu/summary.data.application.html
http://www.kdnuggets.com/datasets/index.html
Clusterwise Regression DataSets