CST/Math/ACSC 209/421-01

Data Mining
Roosevelt University
Syllabus, Fall 2021

Course type

The course is face-to-face course, which means you must attend the class on campus during scheduled time. See attendance policy in this document

Instructor:

Prof. Alexander Wolpert

Contact: awolpert@roosevelt.edu. *Your message must be sent from RU email.* I'll respond to your emails within 24 hours on weekdays and within 48 hours on weekends
Web page: https://www.roosevelt.edu/academics/faculty/profile?ID=awolpert
Virtual Office hours: Tu, 1:00pm-2:00pm, room AUD 336A. By appointment: please email to me no later than M, 7:00pm to request the next day appointment.

Course meetings: Tu, Th 2:00 pm-3:15 pm, AUD 524

Course prerequisites: MATH/ACSC 246 with a min grade of C- and (MATH 217 with a min grade of C- or MATH 347 with a min grade of C- or ACSC 300 with a min grade of C- or ACSC 347 with a min grade of C- or ECON 234 with a min grade of C-)

Course description: Methods of knowledge discovery in massive data, i.e. the study of computer-assisted process of digging through and analyzing enormous data sets and then extracting the 'meaning' of the data by applying mathematical methods. The methods that we study in this course are designed to predict behaviors and future trends based on existing data. Topics include supervised learning (classification), unsupervised learning (clusterization), and techniques for improving data quality.

Required texts:

- Tan, Steinbach, Karpatne, Kumar. Introduction to Data Mining, 2nd Ed., 9780133128901, Pearson, 2018.
- S. Shalev-Schwartz, S. Ben-David. Understanding Machine Learning, CUP, 2014 (free version: http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/copy.html).
- M. Zaki, W. Meira. Data Mining and Analysis, 2 nd edition, CUP 2020 (free version: https://dataminingbook.info/book_html/)
- **Also required***:* R language interpreter and R-studio environment.

RU's learning goals:

- Knowledge of discipline-focused content;
- Effective communication;
- Awareness of social justice and engagement in civic life.

Course learning objectives: In this course students should learn basic and advanced concepts and methods of Data Mining. In particular students should gain knowledge of the following algorithms, methods and concepts:

- Foundations: PAC learning, Agnostic PAC learning, VC dimensions, Linear vs. nonlinear predictors
- Data preparation: PCA and SVD-based dimensionality reduction
- Classification: Decision trees, SVM, Fisher LDA, Naïve Bayes, perceptron and ANN
- Cluster analysis: k-Means, agglomerative methods, density-based methods

Course learning outcomes: Upon successful completion of the course, students should be able to

- Write simple programs in R,
- Apply correctly aforementioned classification algorithms implemented in R to real datasets,
- Perform cluster discovery in real data with R using package-implemented k-means and hierarchical clustering.

Methodology: Lectures, solving problems in class, individual and group HW assignments and programs.

Class reading dates:

Key: no tag on chapters means that these chapters are from Tan, Steinbach, Karpatne, Kumar; ZM tag means that these chapters are from Zaki, Meira book; SSBD tag means that these chapters are from Shalev-Schwartz, Ben-David book. Letter (G) behind chapters means graduate only

Week №	Topics	Reading Assignments (due by the next class)
1	Rules of the course, website, HW and other logistic issues. Data: types of features and their values. Types of Data sets: record, graph, ordered. Samples, distributions, expectation, mode, frequency, percentile, mean, median, range. What is R?	CH 1. 4, CH 2.1.1, CH.9.3.1, ZM: 1.3.1-1.3.4
2	Variance, covariance. PCA intro. PCA as rotation. PCA in R.	ZM 1.3.2, 2.1.2, 2.2, 7.1,7.2; SSBD 23.1; In BB materials: PCA tutorial, R-PCA tutorial

3	Decision tree learning: Recursive partitioning. Splitting measures. GINI, Entropy and error. Learning D-Trees with R.	CH. 3.1-3.3.4, ZM 19.1,19.2 SSBD 17.1 18.1-18.2
4	DT stopping rules. RPART and C4.5. Examples of Decision Trees in R. Multiclass classification. PAC learning; agnostic PAC learning. Agnostic Pac Learning. Graduate (independently): Performance evaluation: holdout, random sub-sampling, cross-validation, bootstrap. Testing significance: interval for accuracy. Compare performance of 2 models.	CH. 3.4, 3.51-3.5.3 SSBD 2.1-2.4,3.1,3.2,4.1 CH 3.6-3.9 (G)
5	Early stopping rules based on independence of split (aka pre-pruning). CHAID trees in R. Review for the midterm 1	CH. 10.4.1, 3.5.4
6	Pruning and tree-raising. C4.5 pruning. Examples of pruning in R.	CH. 3.5.3-3.5.4; SSBD 18.2.2
7	Minimum description length (MDL) principle. MDL pruning in R. Bayesian Classifiers: Bayes theorem and approach to classification. Grad only: Ensemble methods	SSBD 6.1-6.4, 9.1 CH 3.9.2, ZM 18.1 up to, but not including 18.2 CH 4.10.1-4.10.5 (G)
8	Naive Bayesian classifiers. Naïve Bayesian Classifier in R. Maximum likelihood estimators. Linear predictors Grad only: Random forests.	CH. 4.4 SSBD 24.1, 24.2 4.10.6 (G)
9	Linear predictors continued. Perceptron. Fisher LDA. Fisher LDA in R. Grad only: Instance based classifiers	SSBD 9.1, 9.1.1, 9.1.2, 24.3; CH 4.7.1; ZM 20.1 CH 4.3(G)
10	Learnability of infinite-size classes. VC—Dimension, examples. Fundamental Theorem of statistical learning: 2 versions. Midterm review	SSBD 6.1-6.4
11	VC-dimension of half spaces. Support Vector Machines (SVM): classification by decision boundary, maximum margin hyperplanes. Linear SVM: linear decision boundary and margin of a classifier. Learning SVM and associated constrained optimization problem. SVM in R. Graduate (independently): SVM for linearly non-separable data: Soft margin maximization. Non-linear SVM: kernel trick.	SSBD 9.1.3.; 15.1 ZM 21.1, 21.2; CH. 4.9.1-4.9.3 CH. 2.4.7, 4.9.4. (G) SSBD 15.2 intro-15.2.2, 16.1-16.2(G) ZM 5.1-5.3(G)
12	Basic K-means. Choosing initial centroids. Runtime and space. Evaluating clustering quality. K-means continued: Handling	CH 7.1, 7.2.1-7.2.5 ZM 13.1-13.2 SSBD 22.1-22.2

	empty clusters; Pre- and post- processing; Bisecting K-means. K—means in R. Limitations of K-means and how to handle them.	
13	Hierarchical clustering. Basic agglomerative clustering algorithm (AGNES). Proximity definition. Cluster proximity/similarity measures: Min, Max, Average. Agglomerative clustering continued: Mahalanobis distance measure, Ward measure, objective-function based measures. Limitations of measures. Time, space of agglomerative clustering. Divisive clustering example: MST-based algorithm, Diana. DBSCAN Algorithm; Time and Space; Core, border, noise points; When it works, when it doesn't. Graduate (independently). Cluster validity and validation measures: Measures of Interestingness; via correlation; using similarity matrix, Internal measures, e.g. cluster SSE (Sum of Squared Errors); frameworks for validation –statistical; frameworks for validation – internal: Cohesion and separation – graph view; Supervised measures – entropy, purity, precision.	CH 7.3, ZM 14.1-14.2 CH 8.4, 7.4 CH 7.5.1-7.5.6
14	Mixture models. Using Maximum likelihood to evaluate model parameters. Generic Expectation Maximization. EM- clustering algorithm: using MLE to estimate mixture model parameters. Final review	CH 8.2.2 ZM 13.3.
15	Final exam	

Assignments and Exams:

Exams	Dates
Midterm 1	10/7, 1:50 pm – 3:25 pm
Midterm 2	11/11, 1:50 pm – 3:25 pm
Final exam	12/14, 2:00 pm – 5:30 pm

HW assignments	Availability	Submission
HW-1	09/10	09/16
HW-2	09/17	09/23
HW-3	10/08	10/14
HW-4	10/22	10/28
HW-5	10/29	11/04
HW-6	11/19	12/2
HW-7	12/03	**12/8**

Schedule and due dates:

Please note: The instructor may change any aspect of the course, including assignments and due dates, to meet student needs and interests. All syllabus changes will be announced in class and will be communicated to students via RU email and Bb site. Students are responsible for attending class and checking RU email and Bb site for updates.

Assignment feedback dates:

Graded assignments and exams will be returned to students at most 2 weeks from the due/submission dates

Grading:

The course is graded on the basis of accumulated points earned throughout the course.

Grading scale:
A≥93%; A^-≥89%; B^+≥87%; B≥82%; B^-≥79%; C^+≥76%; C≥70%; C^-≥ 65%; D^+≥55%; D≥50%
Note that graduate students will have to do additional HW exercises and will have to answer additional questions on the exams. Standards of grading are different for graduate and undergraduate students. More exactly point distribution is as follows:

Item	Quantity	Points (each)	Subtotal (points)
Midterm Exam	2	70	140
Final	1	100	240
Home Assignments (Undergrad and Grad)	7	6 ÷15	310 (approximately)
Programming Assignments (Undergrad and Grad)	4	5÷15	340 (approximately)
Assignments - Grad only	5	3÷8	UG-340/G-370 (approximately)
Class participation - bonus pts can be earned:			UG-340/G-370
Grand Total (from which grade is counted):			UG-340/G-370

Course expectations for students:

- o Read your RU email at least once a day
- o Access the course Bb site three times a week: on Fridays after 6pm for newly posted lectures and assignments, and on Tue and Th around 12pm for course updates (if any).

Course Policies

Attendance. Class session attendance is required. *Those who come late more than 15 min into the class are considered absent from the class.* But because of continued COVID problems you are allowed to miss 3 meetings without asking permission. Any absence in class beyond 3 allowed absences that is not caused by documented illness or is not explicitly authorized by the instructor at least a day before will cause stiff penalties including dismissal from the class for repeated offenders.

Assignment submission. All assignments must be submitted electronically via BB submission mechanism. Non-programming assignment solutions can be handwritten and scanned, or edited. A complete assignment submission must be uploaded as a single file in pdf, doc(x), xlsx, csv, or txt format (no multiple jpegs are allowed). If necessary files can be zipped (e.g. complete project). Programs must be in .R format.

Late assignments. No late homework accepted.

Extra credit. No extra credit work beyond scheduled by instructor bonus activities will be given and no make-ups for missed work allowed.

I – grade (incomplete).
A grade of incomplete may be given only with the consent of the instructor and appropriate notification to the Office of the Registrar and the instructor's dean or department chair. A student should only receive an Incomplete grade if:
• The student initiates the request for an incomplete grade before the end of the academic term; *and*
• The student is in good standing the course and has completed a majority of the coursework (usually at least 75% of the coursework); *and*
• A medical condition or other serious, non-academic extenuating circumstance (as documented with the Office of the Dean of Students) prevents them from completing a small portion of the coursework required to complete the course prior to the end of the term; *and*
• The required work may be reasonably completed in an agreed-upon timeframe with the faculty member (no later than the end of the next semester, excluding summer); *and*
• The required work does not require the student to retake any portion of the course.

Academic integrity. For the Academic Integrity Policy on issues such as plagiarism, repurposing, cheating and other forms of academic dishonesty please see the University's policies page, which is available at: University Policies Webpage. Additional guidelines for avoiding plagiarism are available on this webpage: Academic Integrity Guide for Students.