University of Puerto Rico, Mayagüez Campus
Department of Industrial Engineering

Course Syllabus

General Information

Course Number:	ININ 6048
Course Title:	**Knowledge discovery in engineering multivariate data**
Credit-Hours:	Three

Course Description

Considerations in multivariate analysis: different data attributes, different data types, regression vs. classification, supervised vs. unsupervised models, model validation, and overfitting. Data integration. Outlier detection. Dimensionality reduction using Lasso regression, Principal Component Analysis, and Partial Least Squares. Supervised learning using logistic regression, linear discriminant analysis, and decision trees. Unsupervised learning using association rules and clustering methods. Use of statistical computer packages for analyzing and predicting engineering multivariate data.

Prerequisites

ININ 4020 – Applied Industrial Statistics or equivalent course on regression modeling

Textbook and References

- Gareth, J.; Witten, D.; Hastie, T.; Tibshirani, R. (2013) **An Introduction to Statistical Learning with Applications in R**, Springer.
- Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Eikström, C.; Wold, S. (2006) **Multi- and Megavariate Data Analysis, Part 1, Basic Principles and Applications, 2nd ed.** Umetrics Academy. [Available for free on books.google.com]
- Hastie, T; Tibshirani, R.; Friedman, J. (2001) The Elements of Statistical Learning. Springer.
- Witten, I.A.; Frank, E.; Hall, M.A (2011) Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed. Morgan Kaufmann. [ISBN: 978-0-12-374856-0]
- Tan, P.; Steinbach, M.; K., V. (2006) Introduction to Data Mining. Addison Wesley.
- Han, J.; Kamber, M.; Pei J. (2012) Data Mining: Concepts and Techniques, 3rd ed. Morgan Kaufmann
- Montgomery, D. C.; Peck, E.A.; Runger, G. C.; Vining, G.G., 2012, Introduction to Linear Regression Analysis, 5th ed. John Wiley and Sons, Inc.
- Walpole, R.E., Myers, R.H., Myers, S.H., and Ye, K., 2011, Probability and Statistics for Engineers and Scientists, 9th ed. Prentice Hall.

Purpose

Development of empirical linear and non-linear model building skills using a variety of tools from multivariate statistics and data mining. Development of skills to identify the model that best represents the natural relationship between a numerical and/or categorical response, and a high-dimensional set of explanatory variables. Special attention is given to data pre-processing, missing value imputation, outlier detection, feature extraction/selection, and model validation. Introduction to unsupervised learning and modeling techniques for multiple response variables

Course Goals

After completing the course, the student should be able to:

- Recognize the fundamental concepts of multivariate analysis: regression vs. classification, supervised vs. unsupervised, outliers, overfitting, different types of attributes, and different data types.
- Select suitable prediction models for different data types and responses.
- Create parsimonious prediction models.
- Evaluate the performance of different learners.
- Use statistical software to analyze data, develop statistical plots, and create empirical models.
- Solve engineering problems in teams.
- Present results concisely using appropriate statistical outputs, written reports, and oral presentations.

Requirements

All students are expected to come to class on time, and prepared; do all assigned readings and related homework; actively participate in class discussions; and satisfy all assessment criteria to receive credit for the course.

Department and Campus Policies

Class attendance: Class attendance is compulsory. The University of Puerto Rico, Mayagüez Campus, reserves the right to deal at any time with individual cases of non-attendance. Professors are expected to record the absences of their students. Frequent absences affect the final grade, and may even result in total loss of credits. Arranging to make up work missed because of legitimate class absence is the responsibility of the student. (Bulletin of Information Undergraduate Studies).

Absence from examinations: Students are required to attend all required examinations. If a student is absent from an examination for a justifiable reason acceptable to the professor, he or she will be given a special examination. Otherwise, he or she will receive a grade of zero of "F" in the examination missed. (Bulletin of Information Undergraduate Studies)

Final examinations: Final written examinations must be given in all courses unless, in the judgment of the Dean, the nature of the subject makes it impracticable. Final examinations scheduled by arrangements must be given during the examination period prescribed in the Academic Calendar, including Saturdays. (See Bulletin of Information Undergraduate Studies).

Partial withdrawals: A student may withdraw from individual courses at any time during the term, but before the deadline established in the University Academic Calendar. (See Bulletin of Information Undergraduate Studies).

Complete withdrawals: A student may completely withdraw from the University of Puerto Rico, Mayagüez Campus, at any time up to the last day of classes. (See Bulletin of Information Undergraduate Studies).

Disabilities: After introducing and identifying himself/herself to the instructor and the institution as a student with disability, the student will receive reasonable accommodations in his/her courses and evaluations. For additional information, contact Services to Students with Disabilities at the office of the Dean of students (Q-019), 787 – 265 – 3862 ó 787 – 832 – 4040 exts. 3250, 3258.

Ethics: Any academic fraud is subject to the disciplinary sanctions described in Chapter VI of the revised General Student Bylaws of the University of Puerto Rico of the Board of Trustees. The professor will follow the norms established in articles 2.6 to 2.14 of the Bylaws. (http://www.uprm.edu/procuraduria/documentos_oficiales .html)

General Topics

Lectures	Topic	Reading
1 – 2	Introduction to topics in multivariate analysis: different attribute and data types, regression vs. classification, supervised vs. unsupervised models, testing/training/cross-validation, and overfitting.	[James] Ch. 1, Ch. 2 *Supplemental:* [Everitt] Ch. 1, [Tan] Sec. 4.1, 4.5
Workshop	Introduction to R.	[James] Sec. 2.3, Sec. 3.6 *Supplemental:* R Modules in eCourses, Appendix An Aide Memoir for R and S-Plus®
3 - 5	Resampling methods. Stacking and one-class classification. Model evaluation.	[James] Ch. 5, Sec. 6.4, Sec. 6.5 *Supplemental:* [Tan] Sec. 4.5, Sec. 5.6 [Witten] Sec. 8.7
6 – 7	Cleaning and integrating data. Outlier detection. Mahalanobis distance.	[James] Sec. *3.3.4-3.3.6* *Supplemental:* [Tan] Sec. 2.2, 2.4.6-2.4.7, Ch. 10
8	Lasso and Ridge regression.	[James] Sec. 6.2 *Supplemental:* [Hastie] 3.4.3
9 – 10	Exam 1	
11 – 13	Projections. Principal component analysis.	[James] Sec. 6.3.1, Sec. 6.6, Sec. 10.2, Sec. [Eriksson] Ch. 2, 3, Sec. 6.7 *Supplemental:* [Everitt] Ch. 3 [Hastie] Sec. 3.4.4
14 – 15	Partial least squares	[James] Sec. 6.3.2, Sec. 6.7 [Eriksson] Ch. 4 *Supplemental:* [Hastie] Sec. 3.4.4
16 – 17	Logistic regression	*Supplemental:* [Montgomery] Sec. 13.1-13.2 [Hastie] Sec. 4.4 [Witten] Sec. 4.6
18 – 19	Linear, quadratic, and regularized discriminant analysis. Canonical correlation analysis.	*Supplemental:* [Everitt] Sec. 7.1-7.2.2, 7.3.2, 8.3 [Hastie] 4.3
20 – 21	Clustering	[James] Sec. 10.3, Sec. 10.5 *Supplemental:* [Everitt] Ch. 6 [Witten] Sec. 6.8
22 – 23	Decision trees	*Supplemental:* [Tan] Ch. 4
24 – 25	Exam 2	
26 – 28	Association rules	[Tan] Sec. 5.1 - 5.2, Ch. 6
29 - 31	Project presentations	

* Lectures are based on 1.5 contact hours.