Principles of Statistical Mining

(STAT-747-01)

Summer 2020

Online Class: Thurs 2:00-4:00pm
Instructor: Prof. Ernest Fokoué
Room: Bldg 14 – Room 2517

Office: Mon 2:00-3:00pm
Phone: **585-643-5549**
eMail: epfeqa@rit.edu

Objectives: The aim of this course is to introduce the entry level graduate student to the fundamental concepts of data mining and machine learning. The course has been deliberately designed to unfold in a hands-on manner, mainly driven by real life examples. It is hoped that a practical approach will allow the student to enjoy a well-paced and effective introduction to the vitally important subject of statistical learning and data mining. Upon completing this course, the student is expected among other things to:

- Become familiar with the basic concepts and tools of modern statistical learning and data mining, and the way data mining is performed with **R**
- Discover the most common techniques and methods of regression and classification (pattern recognition) and the most popular techniques of dimensionality reduction and clustering
- Gain insights into the fundamental aspects of functional relationships between variables, and learn to appreciate and use the subtle differences between predictive optimality and optimal model interpretability
- Develop the ability to perform a complete statistical learning and data mining analysis on moderate to large size data sets
- Apply some of the data mining and machine learning techniques explored on this course to real life data, especially to data you may have generated in your own research

Textbook: There is no required textbook for this course. I will indicate the source of your reading during the first online session and as extra sources as the course unfolds.

This book could be of interest:
An Introduction to Statistical Learning with Applications in R by **Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani**. Springer, 2013. ISBN 978-1-4614-7138-7

Another text of interest is the eBook authored by Julian Faraway:
Practical Regression and ANOVA using R

Quizzes: Four (4) quizzes will be given during this semester namely at the end of weeks 2, 4, 8, 10. These together will comprise 40% of the final grade. All online through myCourses and timed. All quizzes will be in the form of multiple choice questions. Whenever deemed appropriate I will use the online lecture session to provide pointers to the quiz landscape.

Tests/Exams: Midterm Exam at the end of Week 6 (25%) and Final Exam at the end of Week 12 (35%). All online through myCourses and timed. All quizzes will be in the form of multiple choice questions. Whenever deemed appropriate I will use the online lecture session to provide pointers to the exam landscape.

Grade: The final mark will be determined as follows: Tests/Exams (2) – 60%, Quizzes (4) – 40%, [A final average in the interval 90-100% constitutes an "A"; in the interval 80-90%, a "B"; in the interval 70-80%, a "C"; in the interval 60-70%, a "D"; a final average below 60% constitutes an "F". (It's *possible*, although not to be expected, that a "cut-off" could shift slightly due to the way final averages "cluster", but such a shift would not be upward.)] [*Note*: Students will occasionally ask if their grade can be improved by doing a "project". The answer to this question is always the same: "There are no projects in this course; your grade is based solely on on scores that you earn on the announced computer exercises, quizzes, and tests".]

Office hours: Monday from 2:00pm to 3:00pm. Online lectures: Thursday from 2:00pm to 4:00pm. You may also contact me by email to set an appointment in case you can make it to my regular office hours. If you contact me by email, allow 48 hours for me to reply.

Lecture Notes: Lecture notes will be posted on myCourses. It is your responsibility to read both the notes and the additional resources/material from the book or any other indicated source. Since most of you are probably new to most of the concepts explored in this course, I strongly recommend that you spend at least 10 hours of time to deepen your knowledge on every 3 hours of lecture material. Besides the lecture notes, you may occasionally be given additional resources (journal articles, software vignettes, links to websites, etc) deemed useful for your understanding of the material.

Words for the journey: Throughout this semester, I will do my best to help you learn, understand, and master the material of this course. However, no amount of my dedication or willingness to help can serve any consequential purpose unless you demonstrate a strong desire to learn and excel, a firm dedication to regularly study the material, an indomitable determination to overcome all obstacles, a dynamic diligence in seizing all learning opportunities and indeed an unwavering discipline in doing what is required of you. **I wish you all the very best throughout this semester.**

Homework Assignments: I will assign activities for explorations aimed at deepening your understanding, but those will not be graded. Outlines and/or full solutions will be provided to help you check your progress.

Discussion sessions and chat sessions: I will post a discussion topic every week or every other week. I expect every student to participate actively, and to reply to others' posts. This is an excellent way to learn from one another. At the end of the week, I will post a synthesis of the points of view made. Most of the homework answers will be further discussed during the discussion sessions. I will avail myself every Friday at 5:00pm for a weekly chat session through myCourses. I strongly encourage every student to take part in the chat sessions every week. Come prepared with questions.

Datasets: I will strive throughout this course to make it as hands-on and practical as possible. One of the best way to achieve that is to use as many examples as possible, based on as many varieties of datasets as possible. Many websites provide datasets that are ideal from exploring and deeply understanding regression analysis and data mining and machine learning concepts. Amongst others, the following websites might prove useful to you:

- **UC Irvine Data Mining and Machine Learning Repository**
- Most R packages used in this course provide very interesting data sets. Some of those packages include: **datasets**, **car**, **faraway**, **kernlab**, **mlbench**, **MASS**, **class**, **cluster**, **glmnet**, **caret**, **tree**, **wle2**, **leaps**, **lars**, **lasso**, **lasso2**, **LearnBayes**, **mgcv**, **BMS**, **mpf**, **biglm**, **snow**, **foreach**, **doParallel**, **multicore**, **snowfall**

Computing: There are many different software environments for exploring regression analysis. You can use the one you like the most. As for me, I will use only the statistical programming language R. Since you all have to take the required course known as Statistical Software, you will be learning a fair bit of R throughout this semester. I strongly encourage you to download R and start getting acquainted with it. Many useful resources will be provided throughout the semester to help you learn regression using R. The internet is replete with websites that contain videos, slides, pdf files or pieces of code to help people learn R. Among other useful sites, the following might be of interest:

- The R Project website
- R Studio
- Short Reference Card for R
- R Reference card for Regression
- R Reference card for Data Mining
- Texas A & M site for R resources

I strongly recommend that you install R as soon as possible and start getting yourself acquainted with its inner workings.

Tentative Schedule: Below is a tentative schedule attempting to describe how the course will unfold throughout the semester. As the title says, it is tentative and may change (even substantially) during the semester.

Week	Material Covered
1	Visiting the fascinating world of Statistical Machine Learning (7 Wheels of Statistical Machine Learning)- Data (5 Vs and beyond), Approximation/Function Spaces/Models, Learning Criteria, Estimation and Challenges, Refinement/Selection/Regularization/Compression/Aggregation, Prediction, Theory
2	Discovering the Foundational Concepts of Statistical Machine Learning via **k Nearest Neighbors** Learning Machines (Classification and Regression, Performance Assessment like Accuracy, Train and Test Errors, Sensitivity, Specificity, Receiver Operating Characteristics (ROC) curve, Area Under the Curve, Bias Variance Trade-Off on k and Model Complexity)
3	Gathering Important Mathematical, Statistical and Computational Tools and Concepts Essential to Statistical Machine Learning. Fundamentals of **Resampling** Techniques such as **Bootstrapping**, **Subsampling**, and the Ubiquitous **Cross Validation** (CV), Generalized Cross Validation (GCV). Distances, Metrics, Norms, Vector Spaces, Important Inequalities
4	Unveiling Foundational **Unsupervised Learning** Methods (Density Estimation, Dimensional Reduction, Cluster Analysis, Matrix Factorization, featuring kernel density estimation, principal component analysis (PCA), singular value decomposition (SVD), kMeans, Partitioning Around Medoids (PAM), Gaussian Mixture Models (GMM) and Nonnegative Matrix Factorization (NMF), Spectral Clustering)
5	Exploring the Essentials of Model Based **Regression Learning** (Simple Linear Regression, Estimation, Inference, Prediction, Bivariate Linear Regression, Maximum Likelihood, Bayesian Regression, Collinearity, Univariate Polynomial Regression, Nonparametric Regression, Multiple Linear Regression, Multi-collinearity.
6	Extending Regression to **Logistic Regression** for classification and **Generalized Linear Models (GLM)** for more modelling power. Odds, Link functions, Newton Raphson, Fisher Scoring
7	Venturing in the Human Brain with **Neural Networks**. Perceptron, Multilayer Perceptron, Gradient Descent, Backpropagation of the Gradient, Feedforward Neural Networks, **Deep Neural Networks**, Recurrent Neural Networks

8 Tasting the delights and challenges of **Regularization** and **Model Selection**, Variable Selection and Atom Selection. Model Selection Criteria, and Stepwise, and Bayesian Criteria, with Advanced Regularization and Shrinkage Methods (Dropout, Elastic Net), Motivation for Complexity Control, with an introduction to Regularization and Shrinkage Methods, Bias Variance Trade off, Ridge, LASSO and GLMNET)

9 Exploration of the **Classification and Regression Trees** Learning Paradigm (Piecewise Constant Function Estimators, Supervised Recursive Partitioning of the Input Space, Intuitive and Interpretable Learning Paradigm, Impurity, Optimal Split, Pruning Trees, Interpretability)

10 Encountering and Exploring **Support Vector Machines** for Classification and Regression Learning. Large margin learning machines, with the hinge loss and the epsilon insensitive loss, VC dimension, feature space, kernel trick, support vectors, quadratic programming, kernels as similarity measures, Mercer's kernels, positive definite kernels, Reproducing Kernel Hilbert Spaces

11 Harnessing **Ensemble Learning** for classification and regression by aggregation/combination/averaging of many good candidate base learning machines (Bagging, Random Subspace Learning, Boosting, Adaptive Boosting, Gradient Boosting, Random forest

12 Revisiting Classification Learning via Gaussian and Non Gaussian **Bayes Discriminant Analysis** Learning (Linear Discriminant Analysis, Quadratic Discriminant Analysis, Naive Bayes Classification)