Robert AZENCOTT
email: razencot@math.uh.edu ; office : room 608 , PGH building

Fall 2019 MSDS graduate course (Master of Sciences Statistics and Data Science)
Dept of Mathematics, University of Houston

Course number : Math6350
Course Title : Statistical Learning and Data Mining

Summary:
A typical task of Machine Learning is to automatically classify observed "cases" or "individuals" into one of several "classes", on the basis of a fixed and possibly large number of features describing each "case". Machine Learning Algorithms (MLAs) implement computationally intensive algorithmic exploration of large set of observed cases.
In supervised learning, adequate classification of cases is known for many training cases, and the MLA goal is to generate an accurate Automatic Classification of any new case.
In unsupervised learning, no known classification of cases is provided, and the MLA goal is Automatic Clustering, which partitions the set of all cases into disjoint categories (discovered by the MLA).
Numerous MLAs have been developed and applied to images and faces identification, speech understanding, handwriting recognition, texts classification, stock prices anticipation, biomedical data in proteomics and genomics, Web traffic monitoring,etc.
This MSDSfall 2019 course will successively study :
1) Quick Review (Linear Algebra) : multi dimensional vectors, scalar products, matrices, matrix eigenvectors and eigenvalues, matrix diagonalization, positive definite matrices
2) Dimension Reduction for Data Features : Principal Components Analysis (PCA)
3) Automatic Clustering of Data Sets by K-means algorithmics
3) Quick Reviev (Empirical Statistics) : Histograms, Quantiles, Means, Covariance Matrices
4) Computation of Data Features Discriminative Power
5) Automatic Classification by Support Vector Machines (SVMs)
Emphasis will be on concrete algorithmic implementation and testing on actual data sets, as well as on understanding importants concepts.
Pre-requisites :Undergraduate Courses in basic linear algebra and basic descriptive statistics
Homework and Exams : Homework assignments will be mini-projects applied to actual data sets. MiniProjects will all involve computer implementation and tests of Data Mining and Automatic Learning techniques covered in class , and will be expected to use existing dedicated software tools. The students can freely decide to work either in R, or in Matlab, or in Python. These Homework assignment can be prepared by small groups of 1 to 3 students after agreement with the instructor. Mini-Projects Reports will have to be typed (using LaTeX or Word scientific). One midterm exam will be held in class (1h30) without notes, and will be centered on concepts only (no exercises to solve during exams). Final exam will involve an individual take-home projects.

Final grade = 25% final + 15% midterm + 60% homeworks

Reference Book : Reading assignments will be a set of selected chapters extracted from the following reference text
Introduction to Statistical Learning and Data Mining :
authors : James , Witten, Hastie, Tibshirani (This book is freely available online)