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

Fall 2018 MSDS graduate course (Master of Sciences Statistics and Data Science)
Dept of Mathematics, University of Houston

Course number : Math6397-01 (23067)
Course Title : Machine Learning and Data Mining
Tuesday and Thursday 11.30am-1pm; Room : **SW229**

Summary:
A typical task of Machine Learning is to automatically classify observed "cases" or "individuals" into one of several "classes" or "categories", on the basis of a finite number of features describing each "case". Machine Learning Algorithms (MLAs) can generate an automatic classifyer after intensively exploring a large set of observed cases K_n for which one knows the vector D_n of features values, as well as the category CAT_n of K_n. In the past decades, numerous MLAs have been developed and intensively applied to shapes and objects recognition in artificial vision, face identification, speech understanding, handwriting recognition, texts classification and retrieval, etc. MLAs have now been widely extended to the analysis of high dimensional biological data in proteomics and genes interactions networks, as well as to smart mining of massive data sets gathered by monitoring Web traffic or economic and financial activities.
This fall 2018 course will focus on major MLAs derived from the concept of "Positive Definite Kernels". We will study the implementation, performances, and drawbacks of Support Vector Machines (SVMs) for Classification and Prediction, as well as kernel based Automatic Clustering of large sets of "individuals" into distinct "clusters" of strongly similar indivisuals. Emphasis will be on concrete algorithmic implementation and testing on actual data sets, as well as on understanding importants concepts such as Generalization Capacity.
Pre-requisites :These Pre-requisites will be quickly reviewed during the first lectures.
- basic linear algebra : vectors, scalar product of vectors, matrices, products of matrices, eigenvectors and eigenvalues, diagonalization,
- basic descriptive statistics: histograms, quantiles, means, correlations, and covariances

Homework and Exams : Homework assignments will involve several projects applied to actual data sets. Computer implementation and tests of Machine Learning Algorithms taught in class are expected, either in R, or in Matlab or in Python, and will be facilitated by using existing softwares for these algorithms. Projects Reports will have to be typed (using LaTeX or Word scientific). Midterm exam will be held in class (1h30) without notes, and will be centered on concepts only (no exercises to solve in class). Final exam will involve a synthetic take-home project.
Final grade = 30% final + 10% midterm + 60% homeworks
Reference Books : Typed notes on a few key topics will be handed out.
No single textbook. Reading assignments will be a *small* set of selected chapters extracted from the following reference texts
- 2 chapters in Kernel Methods in Computational Biology :
authors: B. Schölkopf, K. Tsuda, J.-P. Vert
- 3 chapters in The Elements of Statistical Learning, Data Mining :
authors : Freedman, Hastie, Tibshirani