COSI 123A – Statistical Machine Learning
Instructor: Dr. Pengyu Hong. Email: hongpeng@brandeis.edu

Description: Statistical machine learning tools are essential to analyzing large-scale noisy data sets, mining patterns/rules, inferring causal relationships, making predictions, and so on. This course focuses on statistical analysis techniques for learning from data (both numerical and categorical) and deals with the issues of designing algorithms, techniques, and systems which automatically improve with experience. It will give students a thorough grounding in the methodologies, technologies, mathematics and algorithms currently needed by research in learning from large-scale example sets. Hands-on training will be provided. Skills learned in this course will be useful for basic and applied research in the fields of bioinformatics, computational neuroscience, data mining, economics, computer vision and pattern recognition, statistical natural language processing, and so on.

Prerequisites: COSI-29A (basic linear algebra, probability, and statistics), differential calculus, basic programming skills. A very useful link: https://en.wikipedia.org/wiki/Matrix_calculus. Matlab (statistics toolbox and optimization toolbox) will be used by the instructor. Students can choose whatever programming languages they want.

Learning Objectives: Students should understand the underlying mathematics of a wide variety of machine learning techniques, especially, the designs of the objective functions and the common optimization method. They should develop an appreciation for what is involved in learning from data, know how to apply those algorithms/techniques or their variances to real data, understand how to evaluate of machine learning techniques, and understand how to perform model selection.


Reference Books:
(a) The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, Jerome Friedman.
(b) Pattern Classification (2nd ed.) by Richard O. Duda, Peter E. Hart and David G. Stork

Course Plan:
• Introduction (Review basic concepts in Probability and Statistics, Information Theory, Matrix and Vector, Model Selection, and Curve Fitting)
• Linear Regression
• Bayesian Decision Theory
  • Two-category case
  • Multi-category case
• Linear Discriminant Analysis
• Principal Component Analysis
• Support Vector Machines
• Maximum-Likelihood Parameter Estimation
  • Gaussian Model
  • Mixture of Gaussian Model
• The EM Algorithm
• Bayesian Parameter Estimation
  • Gaussian Case
  • General Theory
• Markov Chain Monte Carlo
• Bootstrap Inference
• Boosting
• Hidden Markov Models
• Bayesian Networks

Success in this 4-credit hour course is based on the expectation that students will spend a minimum of 9 hours of study time per week in preparation for class (readings, papers, discussion sections, preparation for exams, etc.).

Evaluation: 7 homework assignment (70%). If the student size is small, one of the assignments will be reading + presentation. Final term projects (30%)

Communications: All communications (such as, course announcements, forum discussions, changes of syllabus, slides, snow-day arrangements, and so on) will be on LATTE.

Disabilities: Brandeis seeks to welcome and include all students. If you are a student who needs accommodations as outlined in an accommodations letter, please talk with me and present your letter of accommodation as soon as you can. I want to support you. In order to provide test accommodations, I need the letter more than 48 hours in advance. I want to provide your accommodations, but cannot do so retroactively. If you have questions about documenting a disability or requesting accommodations, please contact Student Accessibility Support (SAS) at 781.736.3470 or access@brandeis.edu.

Academic Integrity: You are expected to be honest in all your academic work. Please consult Brandeis University Rights and Responsibilities for all policies and procedures related to academic integrity. Students may be required to submit work to TurnItIn.com software to verify originality. Allegations of alleged academic dishonesty will be forwarded to the Director of Academic Integrity. Sanctions for academic dishonesty can include failing grades and/or suspension from the university. Citation and research assistance can be found at LTS - Library guides.

Privacy: This class requires the use of tools that may disclose your coursework and identity to parties outside the class. To protect your privacy, you may choose to use a pseudonym/alias rather than your name in submitting such work. You must share the pseudonym/alias with me and any teaching assistants as needed. Alternatively, with prior consultation, you may submit such work directly to me.