Boston University, College of Engineering
Department of Electrical and Computer Engineering

ENG EC500 B1 (Ishwar) – Fall 2015

INTRODUCTION TO LEARNING FROM DATA

Instructor: Prof. Prakash Ishwar (office: PHO 440, tel: 358-3499, e-mail: `pi@bu.edu`)

Classes: TuTh, 4-6 (PHO 205), **Office hours:** W 2:30-3:30pm, F 10:00-11:00am (PHO 440)

Description:
This is an introductory course in statistical learning covering the basic theory, algorithms, and applications. This course will focus on the following major classes of supervised and unsupervised learning problems: classification, regression, density estimation, clustering, and dimensionality reduction. Generative and discriminative data models and associated learning algorithms of parametric and non-parametric varieties will be studied within both frequentist and Bayesian settings in a unified way. A variety of contemporary applications will be explored through homework assignments and a project.

Prerequisites:
Solid foundation in Probability, e.g., EC381 or EK500 or EC505, Linear Algebra, e.g., EK102 or MA142, Multivariate Calculus, e.g., MA225; prior experience with *Matlab*, e.g., EK 127 is important. Good computer programming skills, e.g., EC327, is desirable.

Outline:

- *Introduction*: overview, key concepts and methodology, probability review, optimal decision rules
- *Classification*: Gaussian discriminant analysis (linear and quadratic), naive Bayes and Bayesian naive Bayes, nearest neighbor classifiers, logistic regression, support vector machines (SVMs), kernel trick, multi-class algorithms, decision trees
- *Regression*: linear (ordinary) least squares, robust linear, ridge, Lasso, trees, kernel
- *Density Estimation*: GMM and the EM algorithm, kernel methods
- *Clustering*: k-means/medoids connection to EM for GMMs, spectral clustering, hierarchical clustering
- *Dimensionality Reduction*: PCA, kernel PCA, MDS, Isomap, LLE, Laplacian eigenmaps, Johnson-Lindenstrauss, feature selection

Grading:

35%	Assignments	∼6 computer and ∼4 paper and pen; penalty for late submission; no assignment accepted after solutions released.
40%	Project	Team project involving algorithm development in *Matlab* or *C/C++*; report and presentation required. Details to follow.
25%	Exam	One exam in early Nov (TBD).

Course web site: `http://learn.bu.edu` will contain wealth of information related to the course (lecture slides, handouts, papers, links, etc.) - for registered students only.

Course references: I will not use a formal textbook this year. There is no single textbook that sufficiently covers the material I will introduce in this course, and the feedback on past textbooks was very mixed. Therefore, you will need to rely on lectures and supplementary

material that will be uploaded regularly to the course web site such as lecture slides, reviews, various derivations, proofs, papers, etc. Below is the list of books that can prove useful for various parts of this course should you like to explore. Each book is on reserve at the Science and Engineering Library (max. 2 hour check-out period).

- T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer, 2nd edition - 2009.
- C.M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. Wiley-Interscience, 2nd edition - 2000.

Matlab: Each computer assignment will involve the use of *Matlab* in order to illustrate and compare the main algorithms discussed in the lectures. You are encouraged to use workstations in the SIGNET (PHO307) laboratory. Registered students should contact the lab administrator `enghelp@bu.edu` for an account (if you do not already have one – please check first) and apply for card access to PHO307 (again, only if you do not already have access – please check first) via Zaius at:

`http://www.bu.edu/dbin/eng/zaius/`

PHO307 is available to registered EC500 B1 students on W 9-11am and F 12-2pm. Information about how you can access Matlab remotely can be found at:

`http://collaborate.bu.edu/engit/MatlabRemoteAccess`

Weka: (Waikato Environment for Knowledge Analysis) is a popular suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. Weka is free software available under the GNU General Public License and has been installed in the PHO307 workstations. You can download and install it in your own computer from:

`http://www.cs.waikato.ac.nz/ml/weka/`

Weka contains a number of data analysis visualization tools and a large suite of machine learning algorithm implementations, together with graphical user interfaces for easy access. These can be used for "black-box" validation of your computer assignments and your class project.

Academic conduct: Collaboration is essential for the course project, permitted on homeworks, but illegal in exams. **If there is collaboration in a homework, it must be explicitly acknowledged, and each collaborator must turn in his/her individual analysis and description of results.** The student handbook defines academic misconduct as follows: *"Academic misconduct occurs when a student intentionally misrepresents his or her academic accomplishments or impedes other students' chances of being judged fairly for their academic work. Knowingly allowing others to represent your work as theirs is as serious an offense as submitting another's work as your own."* Please see the student handbook for procedures to be followed should academic misconduct be discovered.