EECS E6690: Topics in data-driven analysis and computation:
Statistical Learning for Biological and Information Systems
Prof. Predrag Jelenkovic

Course description. Ongoing advancements in information systems as well as the emerging revolution in microbiology and medicine are creating a deluge of data, whose mining, inference and prediction will have an enormous economic, social, scientific and medical/therapeutic impact. For example, in biology, microarray technology is creating vast amounts of gene expression data, whose understanding could lead to better diagnostics and potential cure of cancer. Similarly, in information systems, companies like Google, Amazon, Facebook, etc., are facing various problems on massive data sets, e.g., ranking and community detection. This course will cover a variety of fundamental statistical (machine) learning techniques that are suitable for the emerging problems in these application areas, but also applicable in general. The students will gain a comprehensive knowledge in both supervised and unsupervised learning.

Textbooks. The following two books will represent the supporting references for the course. The books are available online:

- Hastie, T., Tibshirani, R. and Friedman, J. The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition. Springer, 2009. Available online at https://web.stanford.edu/~hastie/Papers/ESLII.pdf
- James, G., Witten, D. Hastie, T. and Tibshirani, R. An Introduction to Statistical Learning Springer, 2014. Available online at https://www.statlearning.com. R code can be found at: https://hastie.su.domains/ISLR2/Labs

Prerequisites. Calculus. Some knowledge of probability/statistics and optimization is strongly encouraged, but not required.

Programming. The course uses R language, and the basics of programming in R will be covered in class. (Familiarity with another programming language, e.g., Matlab, will be helpful, but it is not required).

Grading. Homework (20%) + Midterm (35%) + Final Project (45%).

Syllabus

Week 1:

- Course overview and logistics
- Brief review of statistics:
  - Statistical estimates: biased/unbiased, mean variance, standard error
  - Distribution of statistical estimates: Gaussian, Chi, t-distribution
  - Hypothesis testing: t-test, p-value, confidence interval
- Introduction to statistical/machine learning: supervised vs unsupervised
  - Choice of norms and fitting
  - Problem of overfitting
- Linear Regression: formulas for 1-dim, t-test, confidence intervals, example

Week 2:

- Multidimensional linear regression:
    - Matrix formulation and solution
    - Geometric interpretation and properties of the solution
    - Computational complexity: impact for high-dimensional data
    - Dual dot-product (kernel) formulation
    - Goodness of fit/testing: R-squared, distribution of parameters, estimate of variance, more t-testing
    - Introduction to F-distribution and F-test: examples
    - Issues: nonlinearity, collinearity, etc
    - Comparison to non-parametric methods: KNN

Week 3:

- More on dual (kernel) solution and F-test
- RSS, Chi-squared and Cochran's theorem
- Model selection and regularization: best subset, Ridge, LASSO, and their comparison
- Bias-variance tradeoffs
- Analytical formulas for bias-variance tradeoffs for Ridge regression

Week 4:

- Dimensionality reduction: Principal Component Analysis (PCA)
- Basis expansion and polynomial Kernels, solution for Ridge regression
- Testing: validation approaches
    - K-fold cross-validation
    - LOOCV: Leave One Out Cross-Validation: analytical formulas
- Bootstrap method: variance and median estimation, confidence intervals
    - Application to regression

Week 5:

- Classification: categorical data
    - General setting, error measures and their comparison
    - Logistic regression
    - Discriminant classification: linear vs quadratic
    - Comparison of Logistic regression and LDA
    - Naïve Bayes
    - Proof of optimality of Bayes Classifier

Week 6:

- Prediction accuracy vs interpretability
- Tree-based methods: Decision trees
    - Overfitting
    - Tree pruning for regression and classification
    - Bootstrap improvements: bumping and bagging
    - Random forests

Week 7:

- Boosting methods for regression and classification
    - AdaBoost
- Support Vector Machines:
    - Maximum Margin Classifier: geometric interpretation
    - Support Vector Classifier
    - Support Vector Machine: Kernels

Week 8:

- Midterm: in-class and closed book

Week 9:

- Final project outline
- Crash course in convex optimization
- Dual problem to SVM, and SVM classifier
    - Examples of gene expression classification
- Introduction to general Kernel methods

Week 10:

- Reproducing Kernel Hilbert Space Theory
- Representer Theorem
- Generalized Ridge Regression
- Neural Networks vs Kernel methods
- Gradient descent and path of gradient descent
- Stochastic gradient descent (SGD), momentum
- Introduction to Neural Networks
    - Back propagation algorithm

Week 11:

- Unsupervised learning: clustering, community detection, association, ranking
- Clustering:
    - K-means and tree-based methods
    - Examples of cancer gene expression data
- PCA revisited
    - Eigen-decomposition problem: formal proof

Week 12:

- PCA example
- Other dimensionality reduction techniques:
    - Random projections
    - Compressed sensing
- Information retrieval and ranking: Google's PageRank algorithm
- Market basket analysis: Association rules
    - The Apriori Algorithm
- Introduction to social networks
    - Graph representation
    - Community detection = clustering on graphs
        - K-means and tree-methods for graphs
        - Clique-based and density methods

Week 13:

- Dimensionality reduction on graphs
    - Multi-dimensional scaling
    - Block models
    - Spectral clustering

Week 14:

- Final project presentations