Computational Methods in Biomedical Informatics
Course Instructor: Noémie Elhadad (noemie.elhadad@columbia.edu)

Course overview

The course is intended for 1st-year PhD and MA students in the biomedical informatics program. It provides a detailed overview of computational methods for large biomedical and health datasets.

The course is structured as a series of lectures and applied lab sessions. Lectures are interactive and emphasis is put on discussion of algorithms and metrics, interpretation of results, and assumptions and limitations of methods. Lab sessions are conducted in the R framework. Examples of labs are network analysis of public health datasets, clustering of patients according to their clinical characteristics, language modeling of PubMed abstracts, and classification of tumors.

Grading is based on class participation (10%), lab sessions (25%), midterm (25%), and final project (40%).

Syllabus (subject to change)

I. Basics

1. Probabilities refresher (1.5 weeks)
* Discrete random variables, probabilities, joint probabilities, conditional probabilities, Bayes theorem, chain rule, independence, conditional independence
* Expectation and variance
* Maximum likelihood estimation, smoothing
* Standard distributions (e.g., Bernoulli trial and processes, Binomial, Poisson, Normal)

2. Information theory concepts (1.5 weeks)
* Shannon's contributions, the noisy channel model, quantitative definition of information
* Entropy, joint entropy, conditional entropy, mutual information
* Gibb's inequality, KL-divergence
* Log likelihood, cross entropy and perplexity as evaluation methods

3. Linear algebra refresher (0.5 week)
* Vector space, maps, matrices
* Eigenvectors and eigenvalues

4. Markov processes (0.5 week)
* Stochastic processes and Markov properties
* Representation with transition matrices

II. Methods for representing and analyzing large datasets

1. Networks and graphs (2 weeks)
* Real-world examples (information networks, social networks, epidemiological networks, biological networks, etc.), what are networks used for?
* Network properties and types of networks (e.g., small-world, random, scale-free)
* Clustering coefficient, degree distributions
* Node centrality metrics
* Community detection (the Girvan–Newman algorithm)
* PageRank algorithm for information retrieval

2. Vector-space model (1 week)
* Indexing for information retrieval, inverted index
* Weighing schemes for terms along dimensions (e.g., tf*idf)
* Similarity metrics (e.g., cosine, dice, jaccard)
* Evaluation metrics (e.g., precision, recall, F-measure)

3. Bayes networks (1 week)
* Methodology: semantics of Bayes nets, d-separation
* Reasoning over Bayes nets

4. Time Series (0.5 week)
* Piecewise Linear Representation (PLR original and weighted)
* Application to data mining of large datasets of time series
* Limitations of PLR

III. Methods for discovery and prediction

1. Association rules (1 week)
* Hypothesis generation
* A-priori algorithm

2. Clustering (1 week)
* Hierarchical clustering
* K-means
* Evaluation metrics (e.g., purity, task-based)

3. Classification (2 weeks)
* Methodology: training and testing, feature selection
* Naïve Bayes
* Decision Trees
* Support Vector Machine classifiers (if time permits)
* Evaluation metrics (e.g., accuracy, micro-macro F, ROC)

4. Graphical models (2 weeks)

* Methodology: representation, inference, learning
* Directed and undirected models
* Hidden Markov Models

Readings include the following:
- Introduction to Probability. D. Bertsekas and J. Tsitsiklis. 2008. Chapter 1.
- Introductory Statistics with R. P. Dalgaard. 2004. Chapter 1, Appendix A.
- Information Theory, Inference, and Learning Algorithms. D. MacKay. 2003. Chapters 1, 2, 3.
- Speech and Language Processing. D. Jurafsky and J. Martin. 2009. Chapter 6.
- The Structure and Function of Complex Networks. M. Newman. SIAM Review. 2003.
- Networks, Crowds, and Markets: Reasoning About a Highly Connected World. D. Easley, J. Kleinberg. 2010. Chapter 21.
- The PageRank Citation Ranking: Bringing Order to the Web. L. Page and S. Brin. 1999.
- Introduction to InformationRetrieval. C. Manning, P. Raghavan and H. Schutze. 2008. Chapters 17, 21.
- Top 10 Algorithms in Data Mining, Knowledge and Information Systems. X. Wu et al. 2008.
- Decision Trees. In the Handbook of Computational Linguistics and Natural Language Processing. A. Clark, C. Fox and S. Lappin.
- Pattern Recognition and Machine Learning. C. Bishop. 2008. Chapter 7.
- Introduction to Probabilistic Graphical Models. Jordan and Bishop. 2005.