Applied Bayesian Nonparametrics

CSCI 2950-P: Special Topics in Machine Learning, Fall 2011

Bayesian nonparametric (BNP) models define distributions on infinite-dimensional spaces of functions, partitions, or other combinatorial structures. They lead to flexible, data-driven unsupervised learning algorithms, and models whose internal structure continually grows and adapts to new observations. This course surveys state-of-the-art approaches to Bayesian nonparametrics, from its foundations in stochastic processes to the practical tools needed for large-scale computation.

BNP models involve a wide range of stochastic processes, including the Gaussian, Dirichlet, Poisson, Pitman-Yor, beta, gamma, and Levy families. We introduce and discuss various representations of these stochastic processes, including stick-breaking constructions and marginalized prediction rules (e.g., the "Chinese restaurant process" and "Indian buffet process"). Learning and inference is challenging for these models; we discuss sequential Monte Carlo, MCMC, and variational methods. This leads to algorithms for nonparametric clustering, feature induction, and dimensionality reduction, often of data with interesting spatial, temporal, or hierarchical structure. We discuss both the strengths and weaknesses of "infinite" models, in theory and practice.

Course readings will be drawn from statistical journal articles, and machine learning conference papers, describing contemporary and classical BNP research. Overall grades will be based on classroom participation, presentation of some readings, and a final research project. Students who took CSCI 2950-P in the Spring of 2010 may repeat for credit, as the topic has changed.

Prerequisites: Completion of an introductory course in statistical machine learning, such as Brown's *CSCI 1950-F: Introduction to Machine Learning* or *APMA 2610: Recent Applications of Probability and Statistics*. Sufficient comfort with calculus, linear algebra, and probability to read mathematically sophisticated research papers. Programming abilities for course projects.

Administrative Information

Lectures: Tuesdays and Thursdays from 2:30-3:50pm, CIT room 506, 115 Waterman St.

Instructor:

Erik Sudderth (sudderth@cs.brown.edu; 401-863-7660)

Office Hours: Mondays 2:00-3:00pm, Tuesdays 4:00-5:00pm, CIT room 509.

Grading: Class Participation

Because this is a seminar course on advanced topics, participation in class discussion is critical, and will count towards 30% of overall grades. In addition to expecting regular class attendance and engagement in discussions, participation will be evaluated as follows:

1. For each class, the discussion will be divided into three 25-minute segments. Each segment will focus on either a single conference paper, or part of a longer paper. Students should expect to give an overview presentation, and lead discussion, for two of these segments; Prof. Sudderth will lecture for the remainder of the class meeting time.

2. For one of each lecture's assigned readings, all students are expected to submit brief comments about its strengths, its weaknesses, and the questions it raises. Detailed instructions for the

electronic submission process will be provided later. *Comments are due by 8:00am on the day that paper is discussed.* Late comments will not be given credit, but students can skip comments for four readings over the course of the semester without penalty.

Grading: Final Projects

The final project will count towards 70% of overall grades. Of these points, 10% will be based on a 1-2 page project proposal, due in late October; 20% will be based on a short oral presentation given in early December; and 40% will be based on a technical report describing the results. Specific due dates will be announced later. This technical report should be between 8-12 pages long, in the style of top machine learning conferences. Although the results need not be sufficiently novel for publication, the presentation and experimental protocols should be of high quality. Projects which apply BNP models to the student's own research interests are particularly encouraged.

Tentative Syllabus

Gaussian process (GP) regression, classification, dimensionality reduction, density estimation

Dirichlet process (DP) normalized gamma process representation, stick-breaking construction, Chinese restaurant process (CRP)

Infinite mixtures DP mixtures, Pitman-Yor mixtures, clustering, density estimation

Beta process (BP) binary features, infinite factor analysis, Indian buffet process (IBP), stick-breaking constructions, Levy processes and completely random measures

Learning and inference collapsed Gibbs samplers, truncated samplers, slice samplers, sequential Monte Carlo methods, mean field variational methods

Dependence Dependent DPs, hierarchical DPs, nested DPs, kernels & GPs, normalized measures

Markov models infinite Markov models, hidden Markov models, hidden Markov trees, grammars

Hierarchies hierarchical clustering, nested CRPs, coalescents and species sampling

Applications natural language processing, computer vision, bioinformatics

Theory conjugacy and exponential families, asymptotics, inconsistency and pitfalls