From *SIAM News*, Volume 42, Number 2, March 2009

MMDS 2008: Algorithmic and Statistical Challenges in Modern Large-Scale Data Analysis, Part II

Part I of this article appeared in the January/February issue of SIAM News.

By Michael W. Mahoney, Lek-Heng Lim, and Gunnar E. Carlsson

Algorithmic Approaches to Networked Data

In an algorithmic perspective on improved models for data, Milena Mihail of the Georgia Institute of Technology began by describing the recent development of a rich theory of power-law random graphs, i.e., graphs that are random conditioned on a specified input power-law degree distribution. With the increasingly wide range of large-scale social and information networks now available, however, generative models that are structurally or syntactically more flexible have become necessary. Mihail described two such extensions: one in which semantics on nodes is modeled by a feature vector, with edges added between nodes based on their semantic proximity, and another in which the phenomenon of associativity/disassociativity is modeled by fixing the probability that nodes of a given degree d_i tend to link to nodes of degree d_j.

A small extension in the parameters of a generative model, of course, can lead to a large increase in the observed properties of generated graphs. This observation raises interesting statistical questions about model overfitting, and argues for more refined and systematic methods for model parameterization. It also leads to some of the new algorithmic questions that were the topic of Mihail's talk.

Mihail posed the following algorithmic problem for the basic power-law random graph model: Given as input an N-vector specifying a degree sequence, determine whether a graph with that degree sequence exists and, if it does, efficiently generate one (perhaps approximately uniformly randomly from the ensemble of such graphs). Such realizability problems have a long history in graph theory and theoretical computer science. Because their solutions are intimately related to the theory of graph matchings, many generalizations of the basic problem can be addressed in a strict theoretical framework. For example, motivated by associative/disassociative networks, Mihail described recent progress on the joint-degree matrix realization problem: Given a partition of the node set into classes of vertices of the same degree, a vector specifying the degree of each class, and a matrix specifying the number of edges between any two classes, determine whether such a graph exists and, if it does, construct one. She also described extensions of this basic problem to connected graphs, to finding minimum cost realizations, and to finding a random graph satisfying those basic constraints.

The Geometric Perspective: Qualitative Analysis of Data

Gunnar Carlsson of Stanford University offered an overview of geometric and topological approaches to data analysis, which seek to provide insight into data by imposing a geometry on it. In certain applications, such as physics, the phenomena studied support clean explanatory theories that indicate exactly what metric to use to measure the distance between pairs of data points. This is not the case in most MMDS applications. It is not obvious, for instance, that the Euclidean distance between DNA expression profiles in high-throughput microarray experiments captures a meaningful notion of distance between genes. Similarly, despite the natural geodesic distance associated with any graph, the sparsity and noise properties of social and information networks mean that in practice this is not a particularly robust notion of distance.

Part of the problem is thus to define useful metrics—especially because certain applications, including clustering, classification, and regression, often depend sensitively on the choice of metric. Recently, two design goals have emerged. First, don't trust large distances; because distances are often constructed from a similarity measure, small distances reliably represent similarity, but large distances make little sense. Second, trust small distances only a bit—after all, similarity measurements are still very noisy. These ideas are the basis for much work on Laplacian-based nonlinear dimension reduction, i.e., manifold-based methods currently popular in harmonic analysis and machine learning. More generally, the ideas suggest the design of analysis tools that are robust to stretching and shrinking of the underlying metric, particularly in applications like visualization, in which qualitative properties, such as how data is organized on a large scale, are of interest.

Much of Carlsson's tutorial was devoted to these analysis tools and their application to natural image statistics and data visualization. Homology, the crudest measure of topological properties, captures such information as the number of connected components or the presence of holes of various dimensions in the data. Importantly, although the computation of homology is not feasible for general topological spaces, the space can often be modeled in terms of simplicial complexes, in which case the computation of homology boils down to the linear algebraic computation of the Smith normal form of certain data-dependent matrices.

Carlsson also described persistent homology, an extension of the basic idea in which some parameters, such as the number of nearest neighbors and error parameters, can be varied. A "bar code signature" can then be associated with the data set. Long segments in the bar code indicate the presence of a homology class that persists over a long range of parameter values. This can often be interpreted as corresponding to large-scale geometric features in the data; shorter segments can be interpreted as noise.

Statistical and Machine Learning Perspectives

Given a set of measured values of attributes of an object $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, the basic predictive or machine learning problem is to predict or estimate the unknown value of another attribute y. The quantity y is the "output" or "response" variable, and $\{x_1, x_2, \ldots, x_n\}$ are the "input" or "predictor" variables. In regression problems, y is a real number; in classification problems, y is a member of a discrete set of unorderable categorical values (such as class labels). In either case, this can be viewed as a function estimation problem—the prediction takes the form of a function $\hat{y} = F(\mathbf{x})$ that maps a point $\mathbf{x}$ in the space of all joint values of the predictor variables to a point $\hat{y}$ in the space of response variables, and the goal is to produce an $F(\cdot)$ that minimizes a loss criterion.

Jerome Friedman of Stanford University opened the tutorial "Fast Sparse Regression and Classification" by pointing out that it is common to assume a linear model, in which $F(\mathbf{x}) = \sum_{j=1}^{n} a_j x_j$ is modeled as a linear combination of the n basis functions. Unless the number of observations is much, much larger than n, however, empirical estimates of the loss function exhibit high variance. To make the estimates more regular, one typically considers a constrained or penalized optimization problem

$$\hat{\mathbf{a}}(\lambda) = \operatorname{argmin}_{\mathbf{a}} \hat{L}(\mathbf{a}) + \lambda P_{\gamma}(\mathbf{a}),$$

where $\hat{L}(\cdot)$ is the empirical loss and $P_{\gamma}(\cdot)$ is a penalty term. The choice of an appropriate value for the regularization parameter λ is a classic model selection problem, for which cross validation can be used. The choice for the penalty depends on what is known or assumed about the problem at hand. A common choice is $P_{\gamma}(\mathbf{a}) = \|\mathbf{a}\|_{\gamma}^{\gamma} = \sum_{j=1}^{n} |a_j|^{\gamma}$. This interpolates between the subset selection problem ($\gamma = 0$) and ridge regression ($\gamma = 2$) and includes the well-studied lasso ($\gamma = 1$). For $\gamma \leq 1$, sparse solutions (which are of interest due to parsimony and interpretability) are obtained, and for $\gamma \geq 1$, the penalty is convex.

Choice of an optimal (λ,γ) by cross validation, although possible, can be prohibitively expensive even when the loss and penalty are convex, because of the need to perform computations at a large number of discretized pairs. Path-seeking methods have been studied for this situation. Consider the path of optimal solutions $\{\hat{\mathbf{a}}(\lambda) : 0 \leq \lambda \leq \infty\}$, which is a one-dimensional curve in the parameter space $\mathbb{R}^n$. If the loss function is quadratic and the penalty function is piecewise linear, as with the lasso, then the path of optimal solutions is piecewise linear and homotopy methods can be used to generate the full path in time not much greater than that needed to fit a single model at a single parameter value. Friedman described a generalized path-seeking algorithm that solves this problem for a much wider range of loss and penalty functions (including some nonconvex functions) very efficiently.

In a tutorial titled "Kernel-based Contrast Functions for Sufficient Dimension Reduc-tion," Michael Jordan of the University of California, Berkeley, considered the dimension reduction problem in a supervised learning setting. Such methods as principal components analysis, Johnson–Lindenstrauss techniques, and recently developed Laplacian-based nonlinear methods are often used but have limited applicability—for example, the axes of maximal discrimination between the two classes may not align well with the axes of maximum variance. The hope is that there exists a low-dimensional subspace S of the input space X that can be found efficiently and that retains the statistical relation between X and the response space Y. Conventional approaches to this problem of sufficient dimension reduction make strong modeling assumptions about the distribution of the covariate X and/or the response Y. Jordan considered a semiparametric formulation, in which the conditional distribution $p(Y \mid X)$ is treated nonparametrically and the goal is to estimate the parameter S. He showed that this problem can be formulated in terms of conditional independence and evaluated in terms of operators on reproducing kernel Hilbert spaces (RKHSs).

Claims about the independence of two random variables can be reduced to claims about correlations between them when transformations of the random variables are considered: that is, X_1 and X_2 are independent if and only if

$$\max_{h_1,h_2 \in \mathcal{H}} \operatorname{Corr}(h_1(X_1), h_2(X_2)) = 0$$

for a suitably rich function space $\mathcal{H}$. If $\mathcal{H}$ is L_2 and thus contains the Fourier basis, this reduces to a well-known fact about characteristic functions. More interesting from a computational perspective—given that by the "reproducing" property, function evaluation in an RKHS reduces to an inner product—this also holds for suitably rich RKHSs. This use of RKHS ideas to solve the sufficient dimension reduction problem cannot be viewed as a kernelization of an underlying linear algorithm, as is typically the case when such ideas are used (e.g., with support vector machines) to provide basis expansions for regression and classification. Rather, this is an example of how RKHS ideas provide algorithmically efficient machinery for optimizing a much wider range of statistical functionals of interest.

Conclusions and Future Directions

Along with the topics presented in the tutorials, participants heard about a wide variety of data applications: to movie and product recommendations, predictive indexing for fast Web search, pathway analysis in biomolecular folding, functional MRI, high-resolution terrain analysis, galaxy classification, and other applications in computational geometry, computer graphics, computer vision, and manifold learning. We heard about a novel use of approximation algorithms to probe the community structure of large social and information networks as a way to test the claim that such data are even consistent with the manifold hypothesis (which they clearly are not). In all these cases, scalability was a central issue—motivating discussion of external memory algorithms, novel computational paradigms like MapReduce, and communication-efficient linear algebra algorithms. Interested readers are invited to visit the conference Web site, http://mmds.stanford.edu.

The feedback we received made it clear that MMDS has struck a strong interdisciplinary chord. Thinking ahead to a future MMDS workshop, nearly every statistician we spoke with hoped to see more statisticians; nearly every researcher in scientific computing hoped for more data-intensive scientific computation. Practitioners from application domains called for more applications, and just about every theoretical computer scientist expressed the hope for more of the same. MMDS is generating interest as a developing interdisciplinary research area at the interface between computer science, statistics, applied mathematics, and scientific and Internet data applications.

Keep an eye out for future MMDS workshops!

Michael Mahoney (mmahoney@cs.stanford.edu) is a research scientist in the Department of Mathematics at Stanford University. Lek-Heng Lim (lekheng@math.berkeley.edu) is a Charles Morrey Assistant Professor in the Department of Mathematics at the University of California, Berkeley. Gunnar Carlsson (gunnar@math.stanford.edu) is a professor in the Department of Mathematics at Stanford University.