Meeting: Algorithms for Modern Massive Data Sets

Algorithmic and statistical challenges in modern large-scale data analysis were the focus of MMDS2008. Michael W. Mahoney (Stanford), Lek-Heng Lim (Berkeley) and Gunnar E. Carlsson (Stanford) report: The 2008 Workshop on Algorithms for Modern Massive Data Sets (MMDS2008) was held at Stanford University, June 25–28. Its goals were twofold: first, to explore novel techniques for modeling and analyzing massive, high-dimensional, and nonlinearly-structured scientific and internet data sets; and second, to bring together computer scientists, statisticians, mathematicians, and data analysis practitioners to promote cross-fertilization of ideas. MMDS2008 followed MMDS2006, which was originally motivated by the complementary perspectives brought by the numerical linear algebra and theoretical computer science communities to matrix algorithms in modern informatics applications.

Diverse Approaches to Modern Data Problems

Graph and matrix problems were common topics for discussion, largely since they arise naturally in data mining, machine learning, and pattern recognition. For example, a common way to model a large social or information network is with an interaction graph model, $G = (V,E)$, in which nodes in the vertex set V represent "entities" and the edges in the edge set E represent "interactions" between pairs of entities. Alternatively, these and other data sets can be modeled as matrices, since an $m \times n$ real-valued matrix A provides a natural structure for encoding information about m objects, each of which is described by n features.

It is worth emphasizing the very different perspectives that have historically been brought to such problems. A common view of the data, in particular among computer scientists interested in data mining and knowledge discovery, has been that the data are an accounting or a record of everything that happened in a particular setting. From this perspective, the goal is to tabulate and process the data at hand to find interesting patterns, rules, and associations. A very different view of the data, more common among statisticians, is as of a particular random instantiation of an underlying process describing unobserved patterns in the world. In this case, the goal is to extract information about the world from the noisy or uncertain data that is observed.

Of course, the two perspectives are not incompatible: statistical and probabilistic ideas are central to much of the recent work on developing improved approximation algorithms for matrix problems; much recent work in machine learning draws on ideas from both areas; and in boosting the regularization parameter, i.e., the number of iterations, also serves as the computational parameter.

Given the diversity of possible perspectives, MMDS2008 was loosely organized around six hour-long tutorials that introduced participants to the major themes of the workshop.

Large-Scale Informatics: Problems, Methods, and Models

On the first day of the workshop, participants heard tutorials by Christos Faloutsos of Carnegie Mellon University and Edward Chang of Google Research, in which they presented an overview of tools and applications in modern large-scale data analysis.

Faloutsos began his tutorial on "Graph mining: laws, generators and tools" by motivating the problem of data analysis on graphs. He described a wide range of applications in which graphs arise naturally, and he reminded the audience that large graphs that arise in modern informatics applications have structural properties that are very different from traditional Erdős-Rényi random graphs. Although these structural properties have been studied extensively in recent years and have been used to develop numerous well-publicized models, Faloutsos also described empirically-observed properties that are not reproduced well by existing models. Building on this, Faloutsos spent much of his talk describing several graph mining applications of recent and ongoing interest.

Edward Chang described other developments in web-scale data analysis in his tutorial on "Mining large-scale social networks: challenges and scalable solutions." After reviewing emerging applications—such as social network analysis and personalized information retrieval—Chang covered several other applications in detail. In all these cases, he emphasized that the main performance requirements were "scalability, scalability, scalability."

Modern informatics applications like web search afford easy parallelization, e.g., the overall index can be partitioned such that even a single query can use multiple processors. Moreover, the peak performance of a machine is less important than the price-performance ratio. In this environment, scalability up to petabyte-sized data often means working in a software framework like MapReduce or Hadoop that supports data-intensive distributed computations running on large clusters of hundreds, thousands, or even hundreds of thousands of commodity computers.

Algorithmic Approaches to Networked Data

Milena Mihail of the Georgia Institute of Technology described algorithmic perspectives on developing better models for data in her tutorial "Models and algorithms for complex networks." She noted that in recent years a rich theory of power law random graphs has been developed. With the increasingly wide range of large-scale social and information networks that is available, however, generative models that are structurally or syntactically more flexible are increasingly necessary. By introducing a small extension in the parameters of a generative model, of course, one can observe a large increase in the observed properties of generated graphs.

This observation raises interesting statistical questions about model overfitting, and it argues for more refined and systematic methods of model parameterization. This observation also leads to new algorithmic questions that were the topic of Mihail's talk.

The Geometric Perspective: Qualitative Analysis of Data

A very different perspective was provided by Gunnar Carlsson of Stanford University, who gave an overview of geometric and topological approaches to data analysis in his tutorial "Topology and data." The motivation underlying these approaches is to provide insight into the data by imposing a geometry on it. Part of the problem is thus to define useful metrics—in particular since applications such as clustering, classification and regression often depend sensitively on the choice of metric—and two design goals have recently emerged. First, don't trust large distances: since distances are often constructed from a similarity measure, small distances reliably represent similarity but large distances make little sense. Second, only trust small distances a bit: after all, similarity measurements are still very noisy. These ideas suggest the design of analysis tools that are robust to stretching and shrinking of the underlying metric. Much of Carlsson's tutorial was occupied by describing these analysis tools and their application to natural image statistics and data visualization.

Statistical and Machine Learning Perspectives

Statistical and machine learning perspectives on MMDS were the subject of a pair of tutorials by Jerome Friedman of Stanford University and Michael Jordan of the University of California at Berkeley. 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.

In his tutorial, "Fast sparse regression and classification," Friedman began by noting that it is common to assume a linear model, in which the prediction $\hat{y} = F(\mathbf{x}) = \sum_{j=1}^{n} a_j x_j$. 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. The choice of an appropriate value for the regularization parameter λ is a classic model selection problem. A common choice for the penalty is the ℓ_p-norm of the coefficient vector $\mathbf{a} = (a_1, a_2, \ldots, a_n)$. 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 are obtained, and for $\gamma \geq 1$, the penalty is convex. Although one could choose an optimal (λ, γ) by cross validation, this can be prohibitively expensive. In this case, so-called path seeking methods, that can be used to generate the full path of optimal solutions $\{\hat{\mathbf{a}}(\lambda): 0 \leq \lambda \leq \infty\}$ in time that is not much more than that needed to fit a single model, have been studied. Friedman described a generalized path seeking algorithm, which solves this problem for a much wider range of loss and penalty functions very efficiently.

Jordan, in his tutorial "Kernel-based contrast functions for sufficient dimension reduction," considered the dimensionality reduction problem in a supervised learning setting. Methods such as Principal Components Analysis, Johnson-Lindenstrauss techniques, and Laplacian-based non-linear methods are often used, but their applicability is limited since, e.g., the axes of maximal discrimination between two of the classes may not align well with the axes of maximum variance. One might hope that there exists a low-dimensional subspace of the input space X which can be found efficiently and which retains the statistical relationship between X and the response space Y.

Jordan showed that this problem of Sufficient Dimensionality Reduction (SDR) could be formulated in terms of conditional independence and that it could be evaluated in terms of operators on Reproducing Kernel Hilbert Spaces (RKHSs). Interestingly, this use of RKHS ideas to solve this SDR 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 SVMs) to provide basis expansions for regression and classification. Instead, this is an example of how RKHS ideas provide algorithmically efficient machinery to optimize a much wider range of statistical functionals of interest.

Conclusions and Future Directions

In addition to other algorithmic, mathematical, and statistical talks, participants heard about a wide variety of data applications. Interested readers are invited to see presentations from all speakers at the conference website, http://mmds.stanford.edu.

The feedback we received made it clear that MMDS has struck a strong interdisciplinary chord. For example, nearly every statistician commented on the desire for more statisticians at the next MMDS; nearly every scientific computing researcher told us they wanted more data-intensive scientific computation at the next MMDS; nearly every practitioner from an application domain wanted more applications at the next MMDS; and nearly every theoretical computer scientist said they wanted more of the same.

There is a lot of interest in MMDS 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 MMDSs! ■