

Editorial

A Statistician Thinks About Machine Learning

This short note was prepared for the special issue of Statistica Sinica on Machine Learning and Data Mining.

Since around 1996, when the relationship between large margin classifiers and optimization problems in reproducing kernel Hilbert space became widely known, the line between machine learning as done by computer scientists and (sometimes under another name) by statisticians has become increasingly blurred. Historically, it was typical of computer scientists to work with large data sets and evaluate the methods empirically on set-aside test sets, with little reliance on statistical theory, while statisticians worked with smaller data sets and relied on simulated data and theory for evaluation. Now, statisticians and computer scientists do all of these things.

The papers in this special issue illustrate the growth of this increasingly important area for the field of Statistics, one with implications for the future of the profession. The strong motivation of real world scientific problems for the development of new statistical methods is clear in many of the papers, as is the importance of understanding the scientific context of the data which is the target of newly developed methods. There are other important themes also running through this issue, including model tuning and the estimation of generalization error, selection of relevant subsets of variables or features from among a large number of candidate variables or features, and handling of ever larger data sets by statistically valid compression algorithms and approximation methods.

It is clear from many of the papers that statisticians must develop an ease with numerical methods, approximation theory and mathematical programming, and be able to employ these tools as they develop their data analysis methods. Several papers make it clear that the authors have relied on public software as part of the research behind the papers. It goes without saying, that as models and algorithms become more complex and computer intensive, it behooves authors to produce stable, well

documented public software for others to use their methods and build on their results. Last, but not least, theory has not been forgotten in this interesting collection and retains its traditional role in understanding the properties of new methods.

What would statisticians consider the most important and challenging issues? I can only speak for myself, but we are seeing ever larger data sets in areas that affect peoples' every day lives - large demographic studies and clinical trials, huge collections of diverse data relevant to climate and climate change, and so forth. Researchers in important scientific areas, such as economics, physics, and astronomy, to mention a few, are also seeing larger and larger data sets. We need to be able to learn the science, develop better modeling techniques, sometimes, but not always, specific to the subject area, and we need to be able to present the results in a manner that makes sense to the general public. Open issues certainly include how to analyze the results of "data mining", that is, looking for and finding interesting needles in the haystack, when it is also possible to extract "needles" from noise, especially when multi-level analyses are taking place. With complex new methods we have to establish why we think that the results will "generalize". Here the interplay between the statistician and the scientist in interpreting the results can be very important. Many exciting practical and theoretical issues remain.

— Grace Wahba