

Technical Perspective

Expressive Probabilistic Models and Scalable Method of Moments

By David M. Blei

ACROSS DIVERSE FIELDS, investigators face problems and opportunities involving data. Scientists, scholars, engineers, and other analysts seek new methods to ingest data, extract salient patterns, and then use the results for prediction and understanding. These methods come from machine learning (ML), which is quickly becoming core to modern technological systems, modern scientific workflow, and modern approaches to understanding data.

The classical approach to solving a problem with ML follows the “cookbook” approach, one where the scientist shoehorns her data and problem to match the inputs and outputs of a reliable ML method. This strategy has been successful in many domains—examples include spam filtering, speech recognition, and movie recommendation—but it can only take us so far. The cookbook focuses on prediction at the expense of explanation, and thus values generic and flexible methods. In contrast, many modern ML applications require interpretable methods that both form good predictions and suggest good reasons for them. Further, as data becomes more complex and ML problems become more varied, it becomes more difficult to shoehorn our diverse problems into a simple ML set-up.

An alternative to the cookbook is probabilistic modeling, an approach to ML with roots in Bayesian statistics. Probabilistic modeling gives an expressive language for the researcher to express assumptions about the data and goals in data analysis. It provides a suite of algorithms for computing with data under those assumptions and a framework with which to use the results of that computation. Probabilistic modeling allows researchers to marry their knowledge and their data, developing ML methods tailored to their specific goals.

The following paper is about probabilistic topic models, a class of probabilistic models used to analyze text data. Topic modeling algorithms ingest large


collections of documents and seek to uncover the hidden thematic structures that pervade them. What is special about topic modeling is it uncovers the structure without pre-labeled documents. For example, when applied to a large collection of news articles, a topic-modeling algorithm will discover interpretable topics—represented as patterns of vocabulary words—such as sports, health, or arts. These discovered topics have many applications: summarizing the collection, forming predictions about new documents, extending search engines, organizing an interface into the collection, or augmenting recommendation systems. Topic models have further been adapted to other domains, such as computer vision, user behavior data, and population genetics, and have been extended in many other ways. There is a deluge of unlabeled text data in many fields; topic models have seen wide application in academia and industry.

A topic model assumes a random process by which unknown topics combine to generate documents. When we fit a topic model, we try to discover the particular topics that combined to form an observed collection. I emphasize that a topic model is a special case of a probabilistic model. Generally, probabilistic modeling specifies a random process that uses unobserved variables (such as topics) to generate data; the central algorithmic problem for probabilistic models is to find the hidden quantities that were likely to have generated the observations under study. What makes this problem hard, for topic models and other models, is that the models that accurately express our domain knowledge are complicated and the data sets we want to fit them to are large. The authors developed a new method for fitting topic models and at large scale.

The typical approach to solving the topic-modeling problem is to fit the topics with approximate Bayesian methods or maximum likelihood methods. (The authors here call these “likelihood-

based” methods.) The solution here is different in that the authors use what is called the method of moments. What this means is that they derive average functions of the data that a topic model would generate if it were the true model. They then calculate these average quantities on the observed documents and derive an algorithm to find the particular topics that produce them. Their algorithms scale to large datasets.

The authors prove theoretical guarantees about their algorithm. They make realistic assumptions about text (the “anchor word” assumption of topics) and assume that the data comes from a topic model. They show that, with enough documents, their algorithm—which involves their selection of the quantities to match and the algorithm to match them—finds the topics that generated the data. This is a significant result. Such guarantees have not been proved for likelihood-based methods, like Markov chain Monte Carlo (MCMC), variational Bayes, or variational expectation maximization. More generally, the paper represents an elegant blend of theoretical computer science and probabilistic machine learning.

Finally, I will posit the main question that came to me as I read the paper. The traditional methods in probabilistic machine learning, MCMC and variational Bayes (VB), provide convenient recipes for fitting a wide class of models. In contrast, much of the analysis and mathematical work that goes into method-of-moments solutions is model-specific. Is it possible to generalize method-of-moments for latent variable models so that it is as easy to derive and use as MCMC and VB? Can we generalize to other topic models? How about other graphical models? Are there guidelines for proving theoretical guarantees for other models? 

David M. Blei is a professor of statistics and computer science at Columbia University, New York City, NY, USA.

Copyright held by author.