Tutorial on Probabilistic Topic Models

Prof. Indrajit Bhattacharya

IISc, Bangalore, India

Abstract

Over the last decade, probabilistic topic models have emerged as an extremely powerful and popular tool for analyzing large collections of unstructured data. While originally proposed for textual data, topic models have since been applied for various other types of data, such as images, videos, music, social networks and biological data. In this tutorial, I will discuss both the modeling and algorithmic aspects of topic models. I will review the fundamentals of probabilistic generative models, and explain how they can be applied for textual data, starting from simple unigram models to the Latent Dirichlet Allocation model. Then I will look at the problem of learning and inference using topic models, explain why exact inference is intractable for them, review the principle of inference using sampling, and discuss Gibbs Sampling strategies for inference in topic models. As applications of topic models, we will look at semantic search and sentiment analysis. Finally, I will discuss some short-comings of LDA, and briefly touch upon more advanced topic models, such as syntactic, correlated, dynamic and supervised topic models.

Bio: Indrajit Bhattacharya is an Assistant Professor in the Computer Science and Automation Department at the Indian Institute of Science, Bangalore. His areas of research are Machine Learning and Data Mining, with a focus on Hierarchical Bayesian Models for non-iid data. He completed his PhD in Computer Science from the University of Maryland at College Park in 2006, and his BTech in Computer Science and Engineering from Indian Institute of Technology, Kharagpur in 1999. Prior to joining IISc, he worked as a Research Scientist at the IBM Research Lab, New Delhi.

International Conference on Management of Data
COMAD 2011, Bangalore, India, December 19–21, 2011