VirginiaTech
VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Computer Science Seminar Series, 2012

National Capital Region

Clustering Algorithms for Streaming and Online Settings

Speaker: Prof. Claire Monteleoni
Dept. of Computer Science
George Washington University

Friday, April 27, 2012
1:00PM- 2:00PM, NVC 325

Abstract

Clustering techniques are widely used to summarize large quantities of data (e.g. aggregating similar news stories), however their outputs can be hard to evaluate. While a domain expert could judge the quality of a clustering, having a human in the loop is often impractical. Probabilistic assumptions have been used to analyze clustering algorithms, for example i.i.d. data, or even data generated by a well-separated mixture of Gaussians. Without any distributional assumptions, one can analyze clustering algorithms by formulating some objective function, and proving that a clustering algorithm either optimizes or approximates it. The k-means clustering objective, for Euclidean data, is simple, intuitive, and widely-cited, however it is NP-hard to optimize, and few algorithms approximate it, even in the batch setting (the algorithm known as "k-means" does not have an approximation guarantee). Dasgupta (2008) posed open problems for approximating it on data streams.

In this talk, I will discuss my ongoing work on designing clustering algorithms for streaming and online settings. First I will present a one-pass, streaming clustering algorithm which approximates the k-means objective on finite data streams. This involves analyzing a variant of the k-means++ algorithm, and extending a divide-and-conquer streaming clustering algorithm from the k-medoid objective. Then I will turn to endless data streams, and introduce a family of algorithms for online clustering with experts. We extend algorithms for online learning with experts, to the unsupervised setting, using intermediate k-means costs, instead of prediction errors, to re-weight experts. When the experts are instantiated as k-means approximate (batch) clustering algorithms run on a sliding window of the data stream, we provide novel online approximation bounds that combine regret bounds extended from supervised online learning, with k-means approximation guarantees. Notably, the resulting bounds are with respect to the optimal k-means cost on the entire data stream seen so far, even though the algorithm is online. I will also present encouraging experimental results.

This talk is based on joint work with Nir Ailon, Ragesh Jaiswal, and Anna Choromanska.

Biography

Claire Monteleoni is an assistant professor of Computer Science at George Washington University. Previously, she was research faculty at the Center for Computational Learning Systems, and adjunct faculty in the Department of Computer Science, at Columbia University. She did a postdoc in Computer Science and Engineering at the University of California, San Diego, and completed her PhD and Masters in Computer Science, at MIT. Her research focus is on machine learning algorithms and theory for problems including learning from data streams, learning from raw (unlabeled) data, learning from private data, and Climate Informatics: accelerating discovery in Climate Science with machine learning. Her papers have received several awards, and she currently serves on the Senior Program Committee of the International Conference on Machine Learning, and the Editorial Board of the Machine Learning Journal.