Course Title

On the memory and speed scalability of spectral clustering

Category

Methodology

Target Audience

Graduate students, researchers and professionals working in data science

Prerequisites for Participants

One semester of linear algebra, and some experience in matrix computing on software

Computer and Software Requirements

A laptop with Matlab, Python, or R installed (familiarity with one of them is enough)

Course Description

Abstract

Spectral clustering emerged as a very attractive clustering approach at the beginning of the century. It is simple to implement, can handle nonconvex clusters easily, and has a rich theory. However, it is computationally expensive which has limited its applicability. As of today, there is still continued effort in the machine learning and data mining communities to make it scalable to large data. In this short course we aim to educate a general audience on the mathematical theory, derivation, computing, and application aspects of spectral clustering. In the first part, we will teach some necessary background material: (1) advanced linear algebra, such as matrix norm, SVD, pseudoinverse, and low rank approximation, (2) spectral graph theory, and (3) mathematical formulation and derivation of spectral clustering from three different perspectives (graph cut, random walk, and matrix perturbation). In the second part, we start by presenting two landmark-based scalable spectral clustering algorithms that we have developed based on novel document-term and bipartite graph models. Afterwards, we will report some ongoing work on the memory scalability of spectral clustering by using batch-based learning algorithms. Lastly, we will demonstrate the performance of spectral clustering through applications like image segmentation, document and image clustering. Sample scripts will be provided to the participants so they can test the algorithms on their own during the short course.