Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2020:6**
http://www.math.su.se

Diffusion map - A nonlinear dimensionality reduction method

Thi Thuy Nga Nguyen*

June 2020

Abstract

In this report, we discuss the diffusion map, a nonlinear dimensionality reduction method, which focuses on discovering the underlying geometrical structure of data. We first consider a given dataset as a weighted graph, then construct a Markov chain on this graph. The transition matrix P, expressed by a Gaussian kernel function with a single parameter σ - kernel width, characterizes the local connectivity on the graph. When running the Markov chain forward in time, P^t integrates the local connectivities to describe the global connectivity, leading to the idea of diffusion distance viewed as the average length of all paths connecting two nodes in the weighted graph. The diffusion distance defined from P^t thus contains the information of the intrinsic data geometry. We then use eigen-decomposition of P^t to define diffusion coordinates and the diffusion map. This map reorganizes the data according to the diffusion distance in which the points are close if they are highly connected in the weighted graph. Moreover, the diffusion map embeds the data in a lower-dimensional space where the Euclidean distance is an approximation of the diffusion distance.

Diffusion map is very robust to noise and easy to implement. Nevertheless, it is not trivial to choose the parameters appropriately: the kernel width σ, the timescale t, and the dimension of the embedding space q. While the kernel width relates to how transition probabilities describe local connectivity, the timescale t affects the ability of the diffusion distances to capture global connectivity, and the dimension q characterizes the intrinsic dimension of the representation we would like to discover. All of them are investigated by toy examples in our report.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: tthnga.nguyen@gmail.com. Supervisor: Chun-Biu Li.