University of Colorado at Boulder
Department of Applied Mathematics

Comprehensive Examination

April 29, 2014
1:00 pm
ECOT 226 (APPM Conference Room)

Presenter

Nathan D. Monnig

Title

Extension Problems in Manifold Learning and Network Analysis

Out-of-sample extension problems arise throughout all types of machine learning applications: after computing a function on a discrete dataset, the user must generalize the function to new inputs in the domain. This function could be as simple as a binary classi cation, or more complex such as a general bi-Lipschitz mapping between two metric spaces. In some cases the extension is explicitly provided by the algorithm, as in the case of linear regression. In other cases, the algorithm that computes the mapping for the data provides no such extension procedure, and an approximation must be developed. In this work, we provide solutions to extension problems in manifold learning and the spectral analysis of networks.

Nonlinear dimensionality reduction embeddings computed from datasets typically lack a mechanism to compute the out-of-sample extension. For spectral nonlinear embeddings, a commonly used approximation scheme is the Nystrøm extension. Since these types of mappings are typically only computed for a nite set of data points, the inverse map is also only de ned on these data. In this presentation, we address the problem of computing a stable inverse map to such a general bi-Lipschitz map, as published in [1]. Our approach relies on radial basis functions (RBFs) to interpolate the inverse map everywhere on the low-dimensional image of the forward map. We demonstrate that the scale-free cubic RBF kernel performs better than the Gaussian kernel: it does not su er from ill-conditioning, and does not require the choice of a scale. The proposed construction is shown to be similar to the Nystrøm extension of the eigenvectors of the symmetric normalized graph Laplacian matrix. Based on this observation, we provide a new interpretation of the Nystrøm extension with suggestions for improvement.

Similar out-of-sample problems arise in the analysis of large networks. For example, when querying a portion of an online social network, one is often left with an incomplete adjacency matrix. In particular, many nodes will appear which are connected to in-sample nodes, but for which we do not know the full adjacency list. The Nystrøm extension can be used to extend the dominant eigenvector/eigenvalue pairs from the fully-sampled submatrix to the out-of-sample portion of the adjacency matrix. However, it is known that the eigenvalues are not accurately recovered by this method. As an alternative, we propose a general edge prediction scheme to approximate the full adjacency matrix. Similar to other edge prediction schemes in the literature, the method is based on the number of common neighbors. However, unlike other methods, the algorithm explicitly estimates the probability of an edge between a pair of out-of-sample nodes, based on the observed statistics of the in-sample subnetwork. This exibility allows the algorithm to adapt to various latent network topologies.

References

[1] Monnig N. D., B. Fornberg, and F. G. Meyer, \Inverting nonlinear dimensionality reduction with scale-free radial basis function interpolation," *Applied and Computational Harmonic Analysis*, (2013).