Learning Dynamic Models from Non-sequenced Data

Tzu-Kuo Huang and Jeff Schneider
Machine Learning Department
Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh PA 15213
tzukuoh@cs.cmu.edu

Virtually all methods of learning dynamic models from data start from the same basic assumption: that the learning algorithm will be provided with a single or multiple sequences of data generated from the dynamic model. In this work we consider the case where the data is not sequenced. The learning algorithm is presented a set of data points from the system's operation but with no temporal ordering. The data are simply drawn as individual disconnected points, and usually come from many separate executions of the dynamic system.

Many scientific modeling tasks have exactly this property. Consider the task of learning dynamic models of galaxy or star evolution. The dynamics of these processes are far too slow for us to collect successive data points showing any meaningful changes. However, we do have billions of single data points showing these objects at various stages of their evolution.

At the other end of the spectrum, cellular or molecular biological processes may be too small or too fast to permit collection of trajectories from the system. Often, the measurement techniques are destructive and thus only one data point can be collected from each sample even though a rough indication of the relative timing between samples may be known.

In all of these applications, scientists would like to construct a dynamic model using only non-sequenced, individual data points collected from the system of interest. In this work, we launch investigation of this topic by considering fully observable, continuous-state, discrete-time models. We formalize the task of learning from non-sequenced data under linear models, and discuss several issues in identifiability. We then develop several learning algorithms based on optimizing approximate posterior distributions, and generalize these algorithms to learn nonlinear models through reproducing kernels. These proposed methods essentially carry out expectation maximization, and thus require good initialization. We address this issue by studying a related problem, that of reconstructing a temporal sequence from out-of-order data points. To solve this problem, we propose a convex program that optimizes a measure of temporal smoothness, and develop efficient solution algorithms. We test these methods on several synthetic and real data sets, including a set of gene expression time series data collected from yeast. Experimental results show both effectiveness and limitations of the proposed methods.

Part of this work appeared in [Huang and Schneider, 2009] and [Huang et al., 2010].

References

Tzu-Kuo Huang and Jeff Schneider. Learning linear dynamical systems without sequence information. In *Proceedings of the 26th International Conference on Machine Learning*, pages 425–432, 2009.

Tzu-Kuo Huang, Le Song, and Jeff Schneider. Learning nonlinear dynamic models from non-sequenced data. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, 2010.

Topic: sequence modeling
Preference: oral/poster
Presenting author: Tzu-Kuo Huang