Scientific Data Analysis via Statistical Learning

Raquel Romano
romano at hpcrd dot lbl dot gov
November 2006
Abstract

The emerging field of scientific data mining addresses the need for scientists to search for, quantify, and visualize information from massive data sets generated by large scale observations and simulations. Statistical machine learning algorithms have enormous potential to provide data-driven solutions to fundamental scientific questions. We present results from ongoing projects in astrophysics and climate modeling at LBNL, including the search for Type Ia supernovae from astronomical observations, and the analysis of hurricanes and tropical storms in climate simulations.
Supervised Learning for Supernova Recognition

Supervised learning techniques, also known as classification methods, offer powerful, adaptive methods for the automatic detection and recognition of rare objects from large scale digital sky surveys. We have integrated state-of-the-art classifiers (Support Vector Machines, boosted decision trees) into the nightly pipeline of the Nearby Supernova Factory at Lawrence Berkeley Laboratory, to quickly and automatically rank more than 500,000 subimages found nightly in order of confidence that they contain supernovae.
large sets of astronomical imagery

potential supernova on galaxy

measured features (geometry, brightness, signal-to-noise, ...)

\[
\begin{bmatrix}
11.88 \\
7.847 \\
-0.117 \\
3.269 \\
3.22 \\
10.36 \\
9.692 \\
-0.1523 \\
0.7023 \\
\vdots \\
0.008165 \\
0.05337
\end{bmatrix}
\]

new supernova candidates ranked by confidence

learned decision boundary identifies supernovae
Dimensionality Reduction of Supernova Spectra

Dimensionality reduction techniques decompose data into a small number of basis elements that capture useful attributes of the original signals. Spectra measured from observed supernovae are projected onto a lower dimensional subspace or manifold whose dimensions are associated with physical properties such as luminosity and phase. These decompositions are useful for validating synthetic spectra generated from simulations, classifying observed spectra by luminosity and phase, and discovering new classes of supernovae.
synthetic time-varying spectra

low-dimensional projection encodes parameters such as phase or luminosity

basis spectra

observed time-varying spectra

clustering and matching in low-dimensional space
Spatiotemporal Analysis of Climate Data

Statistical modeling of high-dimensional, time-varying, non-stationary time series enable the discovery of temporal, spatial, and multivariate statistical dependencies in large-scale climate simulations. Using latent variable models of tropical storm trajectories, we can infer the statistical relationships between space-time-varying atmospheric variables along a storm trajectory and the likelihood of the storm evolving into an intense hurricane. Hierarchical, stochastic models can predict the influence of global-scale, long-term climate patterns on such short-term, local events.
collection of tropical storm tracks

multivariate time series for each track

- total precipitable water
- sea level pressure
- surface friction velocity

large-scale, high-resolution simulations yield many examples for probabilistic modeling