United Nations
Educational, Scientific and
Cultural Organization

ICTP - East African Institute
for Fundamental Research
under the auspices of UNESCO

Unsupervised Methods in Machine Learning

Alessandro LAIO
SISSA, Italy

Abstract:

Unsupervised methods in data analysis aim at obtaining a synthetic description of high-dimensional data landscapes, revealing their structure and their salient features.

We will describe an approach for charting complex and heterogeneous data spaces, providing a topography of the high-dimensional probability density from which the data are harvested. We obtain information on the number and the height of the probability peaks, the depth of the “valleys” separating them, the relative location of the peaks and their hierarchical organization. The topography is reconstructed by using an unsupervised variant of Density Peak clustering [1,2] exploiting a non-parametric density estimator[3], which automatically measures the density in the manifold containing the data[4]. Importantly, the density estimator provides an estimate of the error. This is a key feature, which allows distinguishing genuine probability peaks from density fluctuations due to finite sampling. We show that this approach allows identifying the Markov States explored during a protein folding molecular dynamic trajectory directly from the shape of the mulidimensional probability density, namely without exploiting any kinetic information[5].