Megaman: scalable manifold learning in Python. (English) Zbl 1393.68155

Summary: Manifold Learning (ML) is a class of algorithms seeking a low-dimensional non-linear representation of high-dimensional data. Thus, ML algorithms are most applicable to high-dimensional data and require large sample sizes to accurately estimate the manifold. Despite this, most existing manifold learning implementations are not particularly scalable. Here we present a Python package that implements a variety of manifold learning algorithms in a modular and scalable fashion, using fast approximate neighbors searches and fast sparse eigendecompositions. The package incorporates theoretical advances in manifold learning, such as the unbiased Laplacian estimator introduced by R. R. Coifman and S. Lafon [Appl. Comput. Harmon. Anal. 21, No. 1, 5–30 (2006; Zbl 1095.68094)] and the estimation of the embedding distortion by the Riemannian metric method introduced by D. Perrault-Joncas and the second author [“Non-linear dimensionality reduction: Riemannian metric estimation and the problem of geometric discovery”, Preprint, arXiv:1305.7255]. In benchmarks, even on a single-core desktop computer, our code embeds millions of data points in minutes, and takes just 200 minutes to embed the main sample of galaxy spectra from the Sloan Digital Sky Survey – consisting of 0.6 million samples in 3750-dimensions – a task which has not previously been possible.

MSC:
68T05 Learning and adaptive systems in artificial intelligence
62-04 Software, source code, etc. for problems pertaining to statistics
62H30 Classification and discrimination; cluster analysis (statistical aspects)

Keywords:
manifold learning; dimension reduction; Riemannian metric; graph embedding; scalable methods; Python

Software:
Scikit; word2vec; Python; Megaman

Full Text: Link