Matlab Toolbox for Dimensionality Reduction (v0.7.1b)

=====================================================

Information

-------------------------

Author: Laurens van der Maaten

Affiliation: University of California, San Diego / Delft University of Technology

Contact: lvdmaaten@gmail.com

Release date: November 2, 2010

Version: 0.7.2b

Installation

-------------------------

Copy the drtoolbox/ folder into the $MATLAB\_DIR/toolbox directory (where $MATLAB\_DIR indicates your Matlab installation directory). Start Matlab and select 'Set path...' from the File menu. Click the 'Add with subfolders...' button, select the folder $MATLAB\_DIR/toolbox/drtoolbox in the file dialog, and press Open. Subsequently, press the Save button in order to save your changes to the Matlab search path. The toolbox is now installed.

Some of the functions in the toolbox use MEX-files. Precompiled versions of these MEX-files are distributed with this release, but the compiled version for your platform might be missing. In order to compile all MEX-files, type cd([matlabroot '/toolbox/drtoolbox']) in your Matlab prompt, and execute the function MEXALL.

Features

-------------------------

This Matlab toolbox implements 33 techniques for dimensionality reduction and metric learning. These techniques are all available through the COMPUTE\_MAPPING function or trhough the GUI. The following techniques are available:

- Principal Component Analysis ('PCA')

- Linear Discriminant Analysis ('LDA')

- Multidimensional scaling ('MDS')

- Probabilistic PCA ('ProbPCA')

- Factor analysis ('FactorAnalysis')

- Sammon mapping ('Sammon')

- Isomap ('Isomap')

- Landmark Isomap ('LandmarkIsomap')

- Locally Linear Embedding ('LLE')

- Laplacian Eigenmaps ('Laplacian')

- Hessian LLE ('HessianLLE')

- Local Tangent Space Alignment ('LTSA')

- Diffusion maps ('DiffusionMaps')

- Kernel PCA ('KernelPCA')

- Generalized Discriminant Analysis ('KernelLDA')

- Stochastic Neighbor Embedding ('SNE')

- Symmetric Stochastic Neighbor Embedding ('SymSNE')

- t-Distributed Stochastic Neighbor Embedding ('tSNE')

- Neighborhood Preserving Embedding ('NPE')

- Locality Preserving Projection ('LPP')

- Stochastic Proximity Embedding ('SPE')

- Linear Local Tangent Space Alignment ('LLTSA')

- Conformal Eigenmaps ('CCA', implemented as an extension of LLE)

- Maximum Variance Unfolding ('MVU', implemented as an extension of LLE)

- Landmark Maximum Variance Unfolding ('LandmarkMVU')

- Fast Maximum Variance Unfolding ('FastMVU')

- Locally Linear Coordination ('LLC')

- Manifold charting ('ManifoldChart')

- Coordinated Factor Analysis ('CFA')

- Gaussian Process Latent Variable Model ('GPLVM')

- Autoencoders using stack-of-RBMs pretraining ('AutoEncoderRBM')

- Autoencoders using evolutionary optimization ('AutoEncoderEA')

- Neighborhood Components Analysis ('NCA')

- Maximally Collapsing Metric Learning ('MCML')

Furthermore, the toolbox contains 6 techniques for intrinsic dimensionality estimation. These techniques are available through the function INTRINSIC\_DIM. The following techniques are available:

- Eigenvalue-based estimation ('EigValue')

- Maximum Likelihood Estimator ('MLE')

- Estimator based on correlation dimension ('CorrDim')

- Estimator based on nearest neighbor evaluation ('NearNb')

- Estimator based on packing numbers ('PackingNumbers')

- Estimator based on geodesic minimum spanning tree ('GMST')

In addition to these techniques, the toolbox contains functions for prewhitening of data (the function PREWHITEN), exact and estimate out-of-sample extension (the functions OUT\_OF\_SAMPLE and OUT\_OF\_SAMPLE\_EST), and a function that generates toy datasets (the function GENERATE\_DATA).

The graphical user interface of the toolbox is accessible through the DRGUI function.

Usage

-------------------------

Basically, you only need one function: mappedX = compute\_mapping(X, technique, no\_dims);

Try executing the following code:

[X, labels] = generate\_data('helix', 2000);

figure, scatter3(X(:,1), X(:,2), X(:,3), 5, labels); title('Original dataset'), drawnow

no\_dims = round(intrinsic\_dim(X, 'MLE'));

disp(['MLE estimate of intrinsic dimensionality: ' num2str(no\_dims)]);

mappedX = compute\_mapping(X, 'Laplacian', no\_dims, 7);

figure, scatter(mappedX(:,1), mappedX(:,2), 5, labels); title('Result of dimensionality reduction'), drawnow

It will create a helix dataset, estimate the intrinsic dimensionality of the dataset, run Laplacian Eigenmaps on the dataset, and plot the results. All functions in the toolbox can work both on data matrices as on PRTools datasets (http://prtools.org). For more information on the options for dimensionality reduction, type HELP COMPUTE\_MAPPING in your Matlab prompt. Information on the intrinsic dimensionality estimators can be obtained by typing the HELP INTRINSIC\_DIM.

Other functions that are useful are the GENERATE\_DATA function and the OUT\_OF\_SAMPLE and OUT\_OF\_SAMPLE\_EST functions. The GENERATE\_DATA function provides you with a number of artificial datasets to test the techniques. The OUT\_OF\_SAMPLE function allows for out-of-sample extension for the techniques PCA, LDA, LPP, NPE, LLTSA, Kernel PCA, and autoencoders. The OUT\_OF\_SAMPLE\_EST function allows you to perform an out-of-sample extension using an estimation technique, that is generally applicable.

Many of the available functions are also available through the GUI, which can be executed by running the function DRGUI.

Pitfalls

-------------------------

When you run certain code, you might receive an error that a certain file is missing. This is because in some parts of the code, MEX-functions are used. I provide a number of precompiled versions of these MEX-functions in the toolbox. However, the MEX-file for your platform might be missing. To fix this, type in your Matlab:

mexall

Now you have compiled versions of the MEX-files as well. This fix also solves slow execution of the shortest path computations in Isomap.

If you encounter an error considering CSDP while running the FastMVU-algorithm, the binary of CSDP for your platform is missing. If so, please obtain a binary distribution of CSDP from https://projects.coin-or.org/Csdp/ and place it in the drtoolbox/techniques directory. Make sure it has the right name for your platform (csdp.exe for Windows, csdpmac for Mac OS X (PowerPC), csdpmaci for Mac OS X (Intel), and csdplinux for Linux).

Many methods for dimensionality reduction perform spectral analyses of sparse matrices. You might think that eigenanalysis is a well-studied problem that can easily be solved. However, eigenanalysis of large matrices turns out to be tedious. The toolbox allows you to use two different methods for eigenanalysis:

- The original Matlab functions (based on Arnoldi methods)

- The JDQR functions (based on Jacobi-Davidson methods)

For problems up to 10,000 datapoints, we recommend using the 'Matlab' setting. For larger problems, switching to 'JDQR' is often worth trying.

Papers

-------------------------

For more information on the implemented techniques and for a theoretical and empirical comparison, please have a look at the following papers:

- L.J.P. van der Maaten, E.O. Postma, and H.J. van den Herik. Dimensionality Reduction: A Comparative Review. Tilburg University Technical Report, TiCC-TR 2009-005, 2009.

Version history

-------------------------

Version 0.7.1b:

- Small bugfixes.

Version 0.7b:

- Many small bugfixes and speed improvements.

- Added out-of-sample extension for manifold charting.

- Added first version of graphical user interface for the toolbox. The GUI was developed by Maxim Vedenev with the help of Susanth Vemulapalli and Maarten Huybrecht. I made some changes in the initial version of the GUI code.

- Added implementation of Gaussian Process Latent Variable Model (GPLVM).

- Removed Simple PCA as probabilistic PCA is more appropriate.

Version 0.6b:

- Resolved bug in LLE that was introduced with v0.6b.

- Added implementation of t-SNE.

- Resolved small bug in data generation function.

- Improved RBM implementation in autoencoders (note that successful training of an RBM still depends on parameter settings such as weight\_cost and learning rate that can only be set in the train\_rbm.m code).

- Added implementation of Sammon mapping.

- Removed dependency on the Statistics toolbox in Laplacian Eigenmaps.

- Resolved bug in implementation of SPE.

- Various speed and memory improvements by exploiting Matlab's new BSXFUN functionality.

Version 0.5b:

- Resolved issues with unconnected neighborhood graph for LLE and Laplacian Eigenmaps (now works like Isomap).

- Resolved bug in prewhitening of data.

- Improved implementations of SNE and symmetric SNE.

- Resolved two bugs in nearest neighbor intrinsic dimensionality estimator.

- Replaced MDS implementation by implementation for classical MDS.

Version 0.4b:

- Added Symmetric SNE ('SymSNE') implementation.

- Added Landmark MVU ('LandmarkMVU') implementation.

- Added completely new implementation of autoencoders using RBM training.

- Added out-of-sample extensions for (Landmark) Isomap, LLE, Laplacian Eigenmaps, Landmark MVU, and FastMVU.

- Added new 'difficult' dataset to data generation function.

- Improved implementations of NPE, LPP, and LLTSA.

- Resolved issue with parameter parsing in manifold charting.

- Resolved issue with adaptive neighborhood selection combined target dimensionalities higher than 40.

- The number of timesteps t can now be specified in diffusion maps.

- Speed up the implementations of Kernel PCA and Kernel LDA for datasets with more than 3,000 instances (with factor ~5).

- Resolved efficiency issue eigendecomposition performed by diffusion maps.

- Speed improvement in nearest neighbor search for datasets with more than 2,000 datapoints (with assistance from James Monaco).

- Speed improvement of Hessian LLE implementation.

- The toolbox now works without using the Statistics Toolbox.

- Data generation function now also returns the true underlying manifold.

- Resolved issue that might occur when Isomap or FastMVU are employed on a PRTools dataset.

Version 0.3b:

- Improved PCA implementation for cases in which D > N.

- Added implementation of probabilistic PCA (using EM algorithm).

- Added implementation of manifold charting.

- Added function for adaptive neighborhood selection (with assistance from Nathan Mekuz).

- Various speed improvements (with assistance from Nathan Mekuz).

- Added welcome message.

- Added contents information for VER command.

- Fixed issue with divisions by zero in intrinsic dimensionality estimators.

- Removed implementation of ICA from the toolbox.

Version 0.2b:

- Resolved issues in LPP, NPE, LTSA, and Kernel PCA implementations.

- Added implementations of LLTSA and Simple PCA.

- Added Conformal Eigenmaps (CCA) as a postprocessing step for LLE.

- Added MVU as a postprocessing step for LLE.

- Added function for prewhitening of data.

- Added function for precise out-of-sample extensions for PCA, LDA, NPE, LPP, LLTSA, Simple PCA, autoencoders, and Kernel PCA.

- Added six techniques for intrinsic dimension estimation.

Version 0.1b:

- The initial release of the toolbox.

Disclaimer

-------------------------

(C) Laurens van der Maaten, 2010

You are free to use, modify, or redistribute this code in any way you want for non-commercial purposes. If you do so, I would appreciate it if you refer to the original author or refer to one of the papers mentioned above.

Parts of the code were taken from other authors, but often I made numerous improvements and modifications. A list of files in which I use source code from other authors is given below:

- minimize.m: C.E. Rasmussen

- hlle.m, mgs.m: C. Grimes and D. Donoho

- dijk.m: M.G. Kay

- dijkstra.cpp: J. Boyer

- L2\_distance.m: R. Bunschoten

- jdqr.m, jdqz.m: G. Sleijpen

- components.m: J. Gilbert

- hillclimber2c.m, lmvu.m, fastmvu.m, computegr.c, csdp.m, mexCCACollectData2.c, writesdpa.m, sparse\_nn.m, readsol.m, sdecca2.m, hill\_obj.m: K. Weinberger

- llc.m, infermfa.m, mppca.m: Y. Teh

- cca.m, mexCCACollectData.c: F. Sha

- combn.m: J. van der Geest

- sammon.m: G.C. Cawley

- GUI: M. Vedenev

Contact

-------------------------

If you have any bugs, questions, suggestions, or modifications, please contact me:

lvdmaaten@gmail.com