Multiplotlib: A Data Visualization Library and Interface for Multi-View Data

Milind Agarwal
Computer Science
Johns Hopkins University

Abstract

Access to diverse types of data in a dataset makes visualization and modeling more complex but also more revealing. Exploring your data through visualization remains a crucial first step before any subsequent analysis in most research workflows.

However, even with a wealth of visualization libraries like matplotlib and seaborn available to researchers, there is a lack of support for truly integrative data analysis pipelines for datasets with multiple views. Jointly visualizing multi-view data remains a challenge because of the huge number of features and samples in views and the immense wealth of interactions/correlations between them.

There is a wealth of visualization softwares and libraries like matplotlib, seaborn for Python users and ggplot for R users.The visualization libraries that are open-source (matplotlib,seaborn, ggplot) can be modified by the user to support some very rudimentary visualizations to compare data from differ-ent views. But while these existing libraries could be modified to support multi-view data visualizations, they do not actively encourage its use[2]. To enable users and researchers to perform more intricate and in-depth exploratory data analysis there is a need for more sophisticated, useful and robust toolkits for multi-view data visualization.

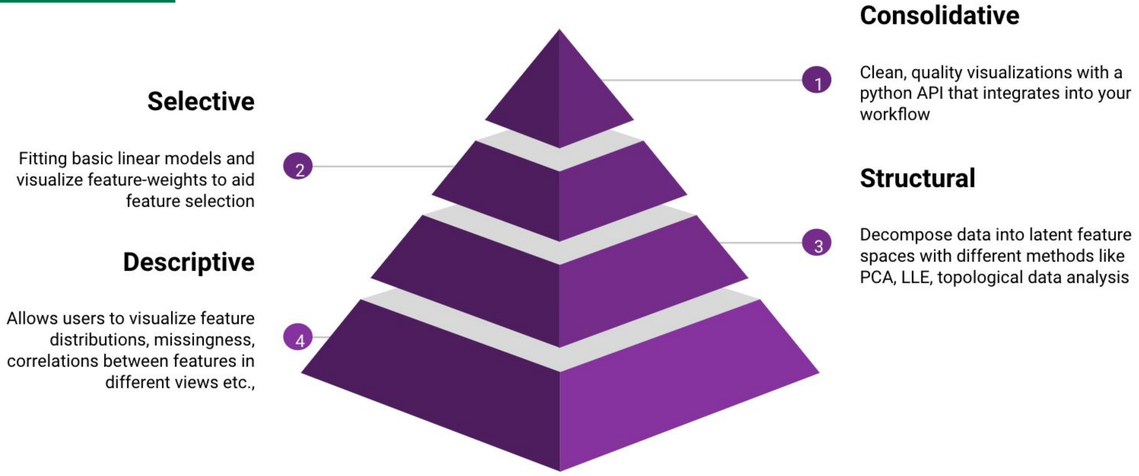
We present multiplotlib, a Python-based visualization library to make statistical graphics. It is built on top of matplotlib and seaborn and is closely integrated with numpy and pandas libraries to offer full flexibility.

Methodology and Approach

Multiplotlib is unique in that it is the first visualization library that will offer in-built support for extensive multiview data analysis for research purposes.

To the best of our knowledge, no other such visualization and analytics library is available for Python. Multiplotlib exposes the following functionalities to users:

- Visualizing useful heuristics necessary for data cleaning e.g. missingness in data views, distributions of features across samples



- Charting correlations among features in a view and across views, corrected by False Discovery Rate methods (e.g. Benjamini Hochberg) or Family-Wise Error Rate methods (e.g. Bonferroni Correction)

- Structurally decomposing views to get the high-level latent features in a view and across views using dimensionality reduction techniques like Principal Component Analysis (PCA), Local Linear Embeddings (LLE) etc.

- Fitting basic linear models (with or without interaction terms) and visualizing feature-weights to aid feature selection during exploratory stage of projects

- Provides a set of visualization and analytical tools to aid researchers in preprocessing and cleaning, exploratory data analysis and modeling.

- Delivers a free and open-source Python visualization library, built on top of popular existing python libraries, designed specifically for datasets with multiple views.

Acknowledgements

This work was funded by the Pistritto Fellowship for Information Visualization Research provided by the Johns Hopkins' Department of Computer Science. I thank my advisor Prof. Alexis Battle and graduate student Benj Shapiro for their valuable advice and support along the way.

BACKGROUND

The rise of big data in the past decade has been a double-edged sword. While access to a wide variety of data for research studies is extremely useful, it can be almost too easy to believe in representations that a subset of our data provides us with.

A single representation of data may generate a misinterpretation of the information[1] as it really only provides us information through one point of view/lens into the intricate problems that we as researchers are trying to solve.

Results

While we were unable to complete this project in its entirety, we plan to make multiplotlib available as an open-source Python library.

This project makes some key contributions to preliminary research in this field:

- Takes a step towards overcoming the shortage of effective statistical and bioinformatic tools[3] for truly integrative multi-view data analysis.

References

[1] Roberts, J. C. (1998). On encouraging multiple views for visualization. Proceedings of the IEEE Symposium on Information Visualization, 8–14. https://doi.org/10.1109/iv.1998.694193

[2] Baldonado, M. Q. W., Woodruff, A., Kuchinsky, A. (2000). Guidelines for using multiple views in information visualization. Proceedings of the Workshop on Advanced Visual Interfaces, 110–119. https://doi.org/10.1145/345513.345271

[3] Shen, R., Olshen, A. B. Ladanyi, M. (2009) Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis.. Bioinformatics 25, 2906– 2912