SYDE 572 Zelek
Fall 2025
Project

Dimensionality Reduction Project

Due: Dec. 1, 2025 midnight

You will also provide a 5-10 minute presentation in class that shows the problem to be solved, the dataset used, and your results and conclusions.
15 marks will be allocated to the submitted report and 5 marks for the presentation.

Reading Material: Chapter 2 and Appendix B4 of book: Vidal, R., Ma, Y. and Sastry, S.S., Generalized Principal Component Analysis. 2016. Interdisciplinary Applied Mathematics, 40.

You may do this individually or as a group of 2.

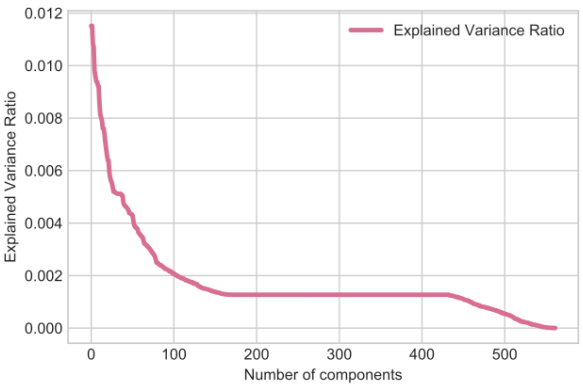
The projects goal is to take a dataset that has at least 10 features and either do classification, linear regression or clustering to solve the problem at hand.
The operation should be done (1) on the original dataset with all the features, (2) a PCA version of the dataset plotting the performance metric against how many features are used (linear dimension reduction), and (3) use ONE non-linear dimensional regression technique (t-SNE, ISOMAP, or UMAP) plotting performance metric against how many features are used in the new space.

The dataset is selected from
https://archive.ics.uci.edu/datasets
with the constraint that the dataset chosen should have more than 10 original features or dimensions.

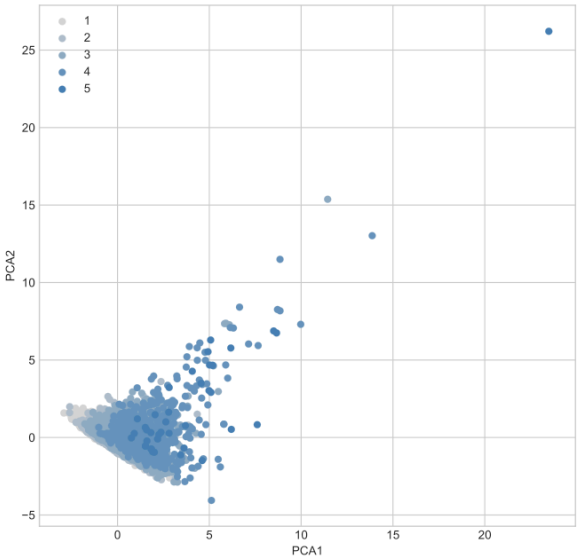
The goal of the project is to run classification, regression or clustering on the original dataset and then apply at least TWO dimensionality reduction algorithms (one linear; and the other non-linear: t-sne, UMAP or ISOMAP) and report any observations. The student is to show the performance metric of the task and how it varies across how many dimensions are chosen in the linear and non-linear reduced space. Also show plots on how the data looks in the new dimensional space.

For the linear dimensional reduction in PCA space, include also the plot that maps the explained variance vs. the number of dimensions used.

For example:



Also, show the distribution of the data across the principal components (probably the major dimensions against each other), as in the following:



In addition, plot the loss or an appropriate performance metric (accuracy?) vs. the number of dimensions used.

For the non-linear case, only plot the data in the new dimensional space and comment on how well the mapping separates the data. Also, plot the performance metric against how many of the new dimensions are used.

For the results, results should be shown on the training dataset as well as on the testing dataset.

What is to be submitted should include a Jupyter file of the code and a pdf version of this. Also include a report submitted in pdf format that describes the following:

1) the dataset chosen, why is it challenging, the problem to be solved for the dataset and what performance metric you will use.
2) plotted results and captions that describe the visuals.
3) Comments about the original results, linear and non-linear dimensionality reduction results.
4) If you can, find some data from elsewhere and test one of your models on this data, do you do as well as your testing data results? This gives some indication of how well your model generalizes.
5) What prevents you from getting perfect results?
6) Any other comments describing what you observed and generalized from the results.