DATA MINING

Dimensionality Reduction(PCA)

Bingan Feng/Wei Zhang/Isabella Wang

Fall 2019

Dimensionality Reduction(PCA)

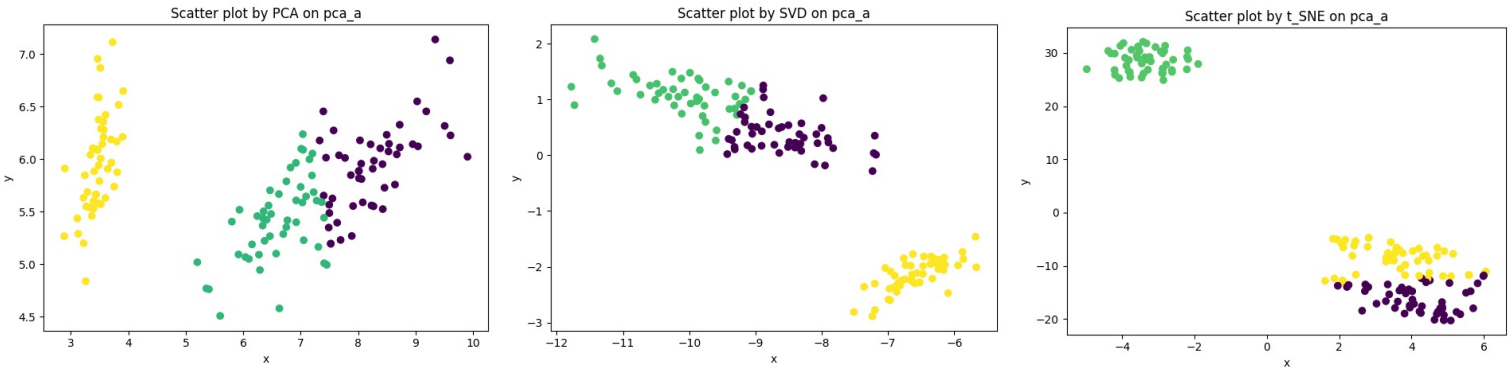
PCA

PCA is a data analysis approach to reduce dimensions of dataset with the minimal lose of informations. Here below are 5 steps to perform a principle component analysis:

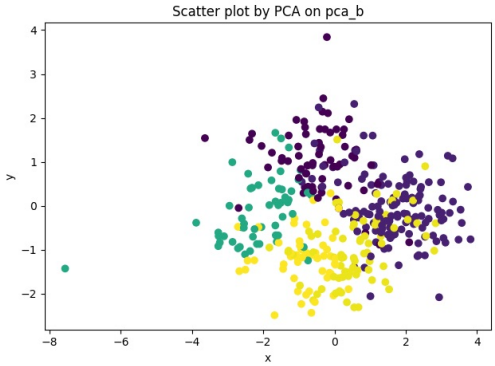
1. Take the whole dataset consisting with d-dimensional samples and ignore the class labels, then store the dataset in form of a $m * d$ matrix where columns are attributes and every row represent a sample ,we denote it as X.
2. Standardize and compute covariance matrix via the following equation:
$$C = \frac{1}{m-1}(X - meanvec)^{\top}(X - meanvec)$$
3. Get the eigenvectors $e_1, e_2, \cdots, e_d$ and eigenvalues $a_1, a_2, \cdots, a_d$ of covariance matrix C.
4. As the project description, select the biggest two Eigenvalues and their corresponding Eigenvectors each Eigenvectors is the column of the $m * 2$ matrix P.
5. In the last step we use following equation to transform our samples onto the 2-dimension subspace: $pc = XP$

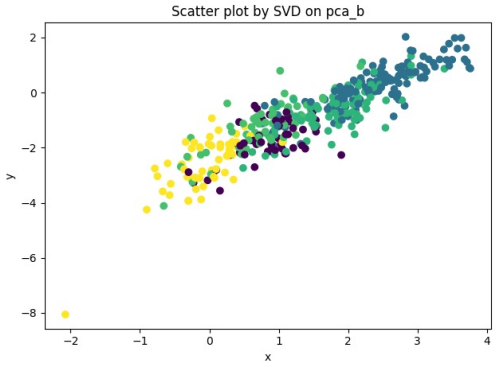
The scatter plots of three datasets are following:

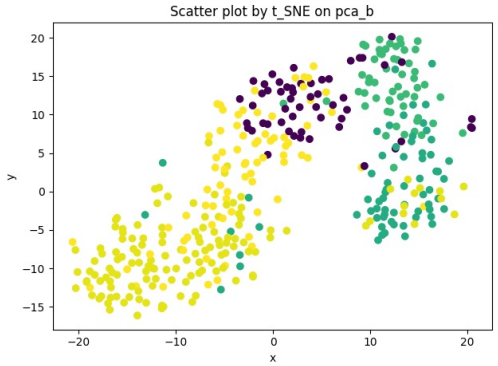
Plot of data_a



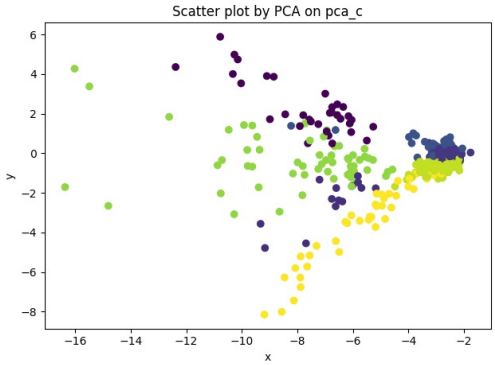
Plot of data_b

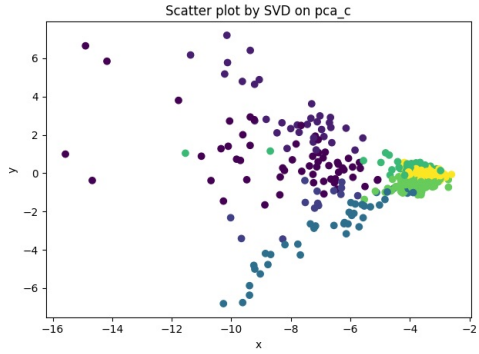


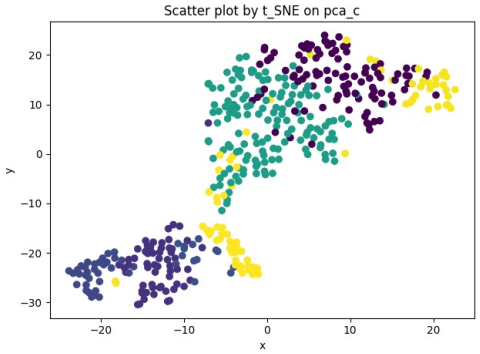




Plot of data_c







Inference

- We observe that t_SNE plot change every execution of the program, it is because t_SNE uses probabilistic approach to reduce dimensions
- The plot of PCA and SVD will be mirror image if we use standardize the data.