CSE 601 – Data Mining and Bioinformatics (Fall 2017)

Project 1:

Dimensionality Reduction & Association Analysis

Part 1: Dimensionality Reduction

Team Members:

1. Siddharth Pateriya spateriy@buffalo.edu Person #: 50206348
2. Hrishikesh Saraf hsaraf@buffalo.edu Person #: 50205927
3. Ajay Gandhi agandhi3@buffalo.edu Person #: 50207403

Part 1: Dimensionality Reduction

Principal Component Analysis:

- PCA stands for Principal Component Analysis, it is the most common form of factor analysis, in which dimensionality reduction takes place in such a way that new dimensions are created which are linear combinations of the original ones.
- It uses eigenvectors and eigenvalues of the data matrix, these eigenvectors have the property that they point along the major directions of variation in the data.
- The main task of PCA is to find principle components of data, by converting a set of correlated variables into a set of linearly uncorrelated values.

- **Steps** for **PCA**:

  1. First we have read the files containing the input data and understood the dimensions of the data.

  2. We assumed m, n to be the data dimensions. Divided the data into values and labels to create an 'mX1' list to hold the labels and 'mXn' matrix to hold the data values.

  3. Next we have calculated the mean matrix of the data, we have done this by calculating the mean across all columns in the data matrix.

  4. Further, we calculated the covariance of the data using the function 'np.cov' (python).

  5. We then calculated the eigen vectors and values from the data matrix using the function 'np.linalg.eig' (python).

  6. Next we created eigen value, eigen vector pairs and sorted them in descending order based on eigen values.

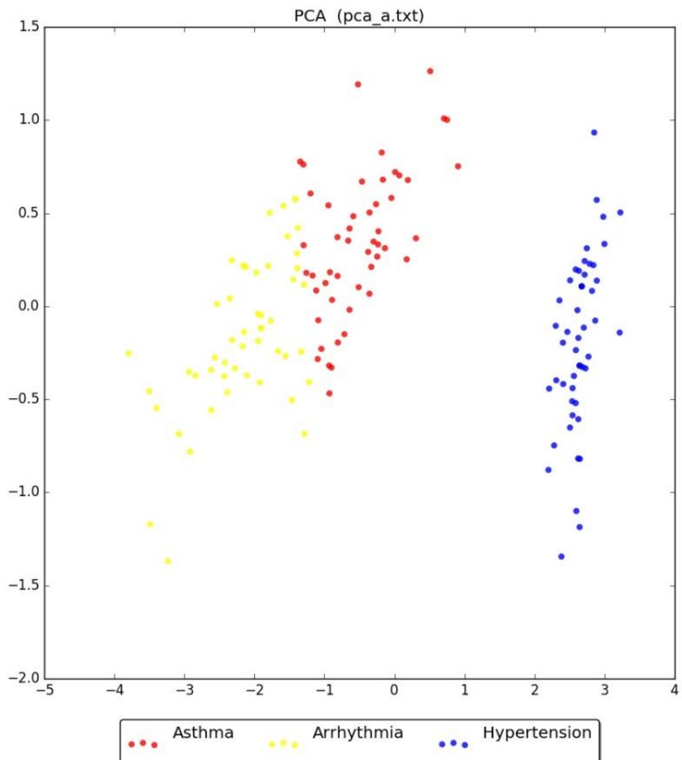
  7. We then selected the top two pairs; which resulted into **dimensionality reduction** and the top two eigen vectors formed the projection matrix.

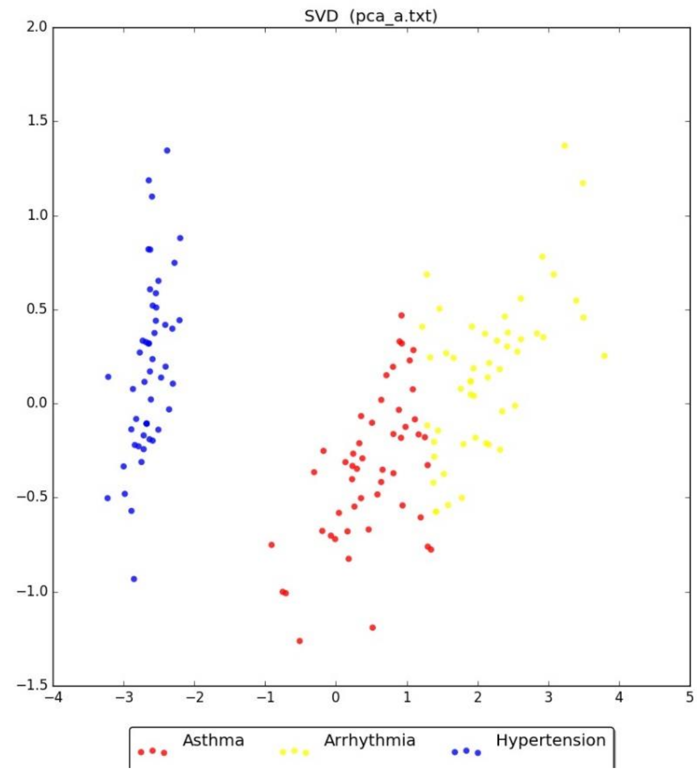
  8. The product of original input data and projection matrix now gave us the new dimensionally reduced data with only two dimensions.

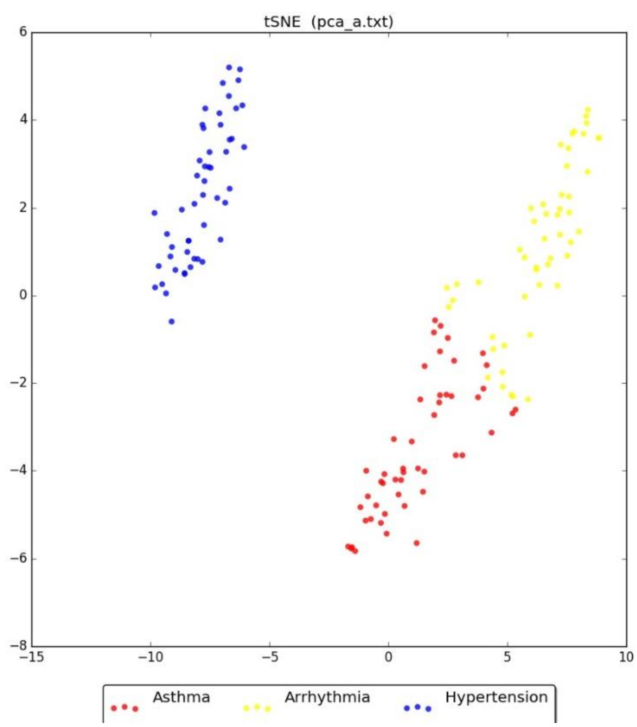
  9. The final step involved generation of a simple scatter plot using library like 'matplotlib' (python).

Output for Dimensionality Reduction:

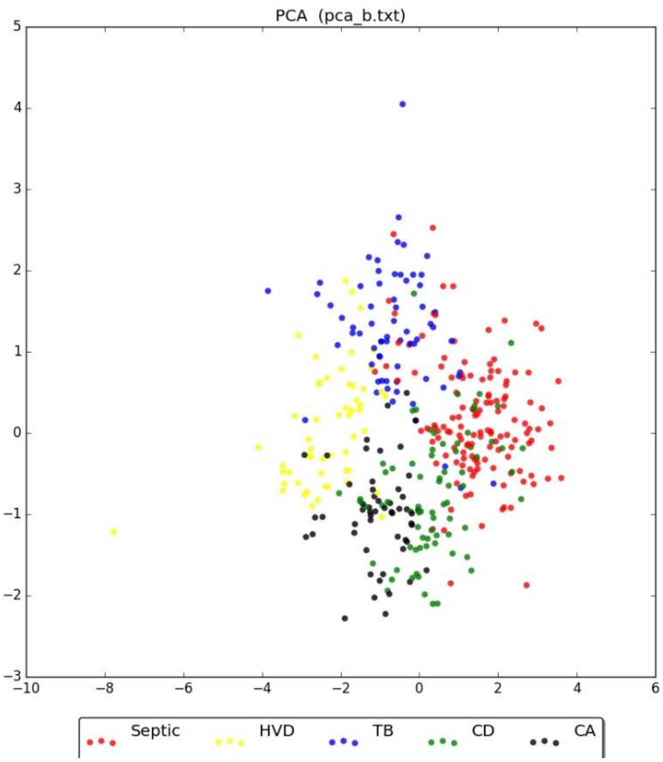
i) Plots for pca_a.txt

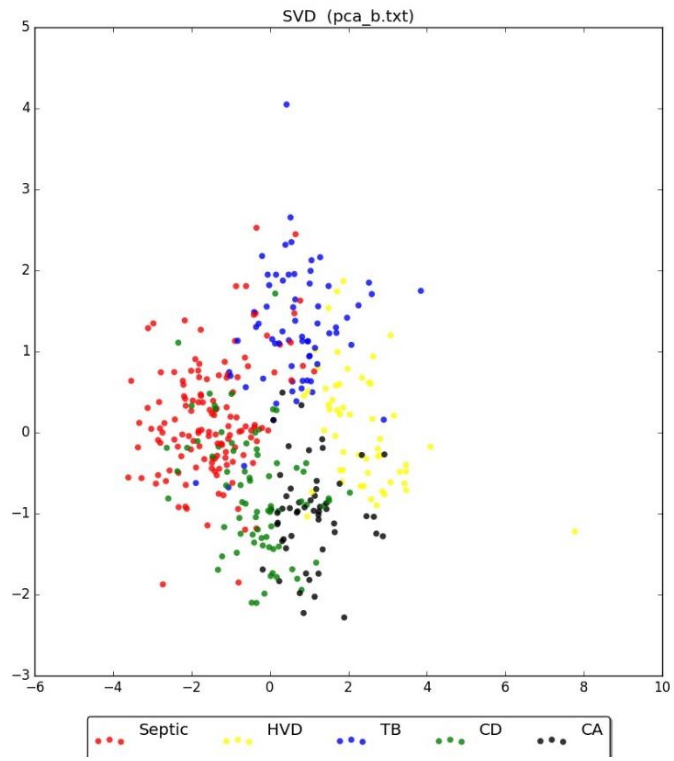


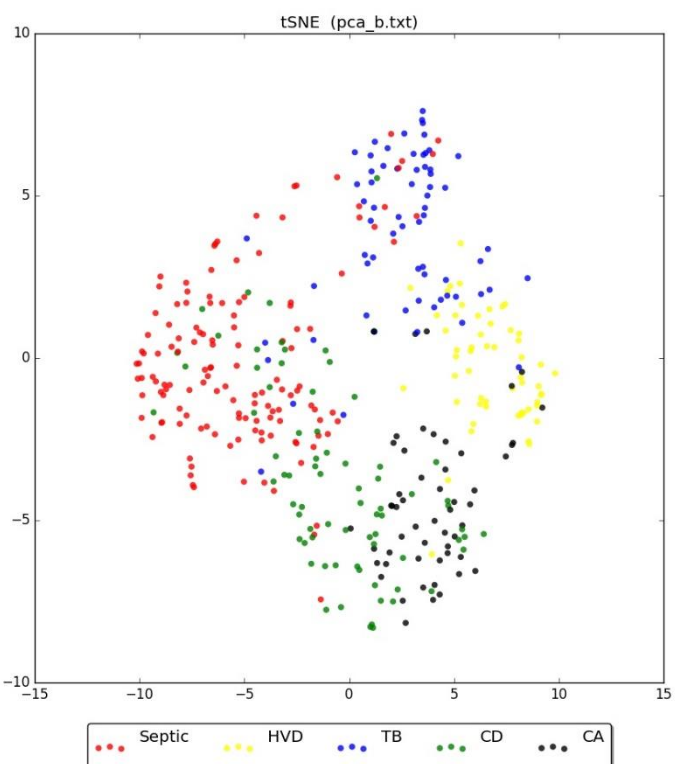




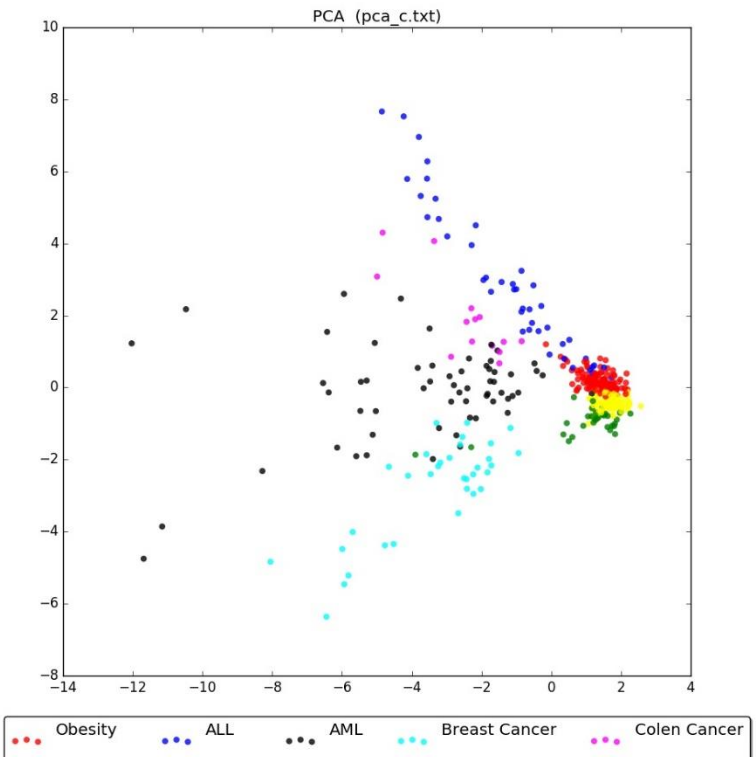
ii) Plots for pca_b.text

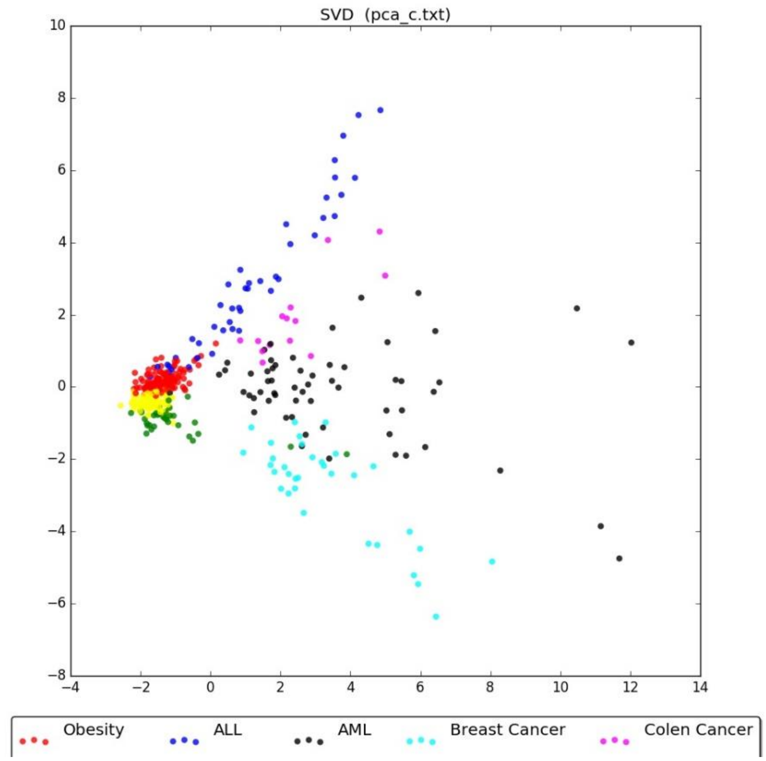


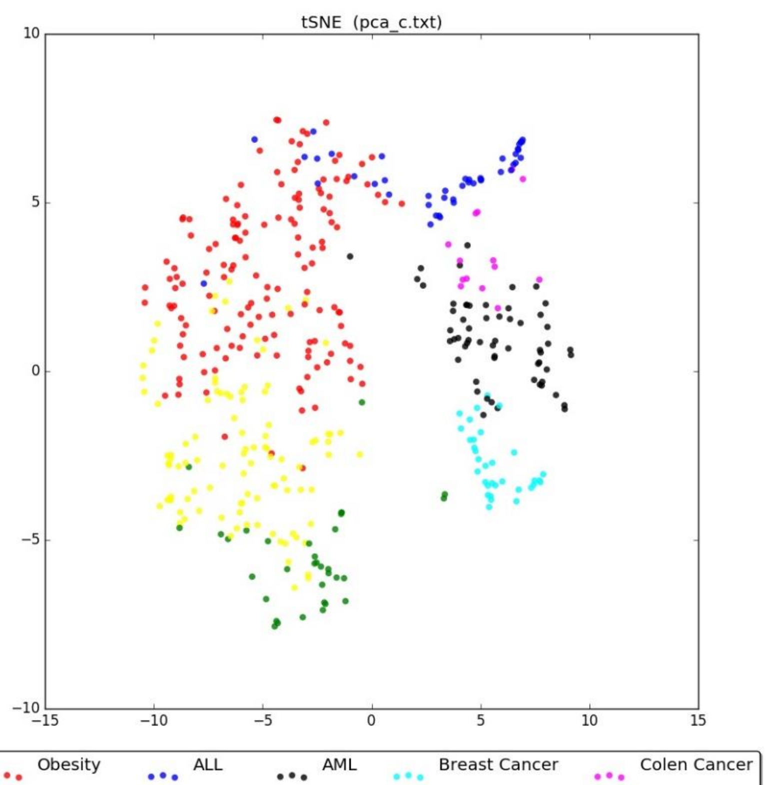




iii) Plots for pca_c.txt







Comparing the results:

- Principal component analysis (PCA) is usually explained via an eigen-decomposition of the covariance matrix. However, it can also be performed via singular value decomposition (SVD) of the data matrix.
- This is why PCA and SVD tend to give similar results in any test case, because the approach is similar i.e. reduce high dimensional data into low dimensional data.
- t-Distributed Stochastic Neighbor Embedding (t-SNE) is also a technique for dimensionality reduction and is mostly suited for the visualization of high-dimensional datasets.
- The drawback however is that in case of high dimensional data, we may need to apply another dimensionality reduction technique as t-SNE leads to huge unnecessary computations and memory consumption.
- We observe that for all pca_a.txt, pca_b.txt and pca_c.txt PCA and SVD tend to give almost identical results on the scatter plot, indicating use of either is fine. However, t-SNE tends to give different results than the PCA and SVD and performs better in case of pca_c.txt and gives more distinction than PCA and SVD.