CSE-601: Data Mining and Bioinformatics

Project 1 Report
Dimension Reduction

Junsong Huang(50292448)
Yi Jin(50295580)
Yin Gong(50290271)

September 24, 2019

1. Absctract

When dealing with industrial level data, these data are often complex and noisy with many attributes that are uncorrelated to our goal. The aim of this project is to apply three kinds of dimension reduction methods naming Principal Component Analysis(PCA), Singular Value Decompostion(SVD) and t-Distributed Stochastic Neighbor Embedding(t-SNE), on give datasets. By reducing the number of dimensions without much loss of information, a certain patterns in the data can be found, which is helpful for distinguishing data and performing better data analysis.

2. PCA

Principal Component Analysis analyzed the data and made linear combination of all attributes to create vectors which were ranked by the magnitude of variances according to the vectors' direction. The newly generated vector with the largest variance is called the first principal component. By applying the vector components on raw data, the main components of the data were kept.

The steps of PCA implementation

1. Read the datasets
2. Extract the data without label and perform normalization by subtracting the mean for each attribute
3. Apply np.cov() and np.linalg.eig() to obtain eigen vector and eigen value of the modified data set
4. Choose two eigen vectors with largest eigen values
5. Perform matrix multiplication on eigen vectors and raw datasets
6. Plot the PCA dimension reduction result on 2-D graph.

3. Comparison between three methods

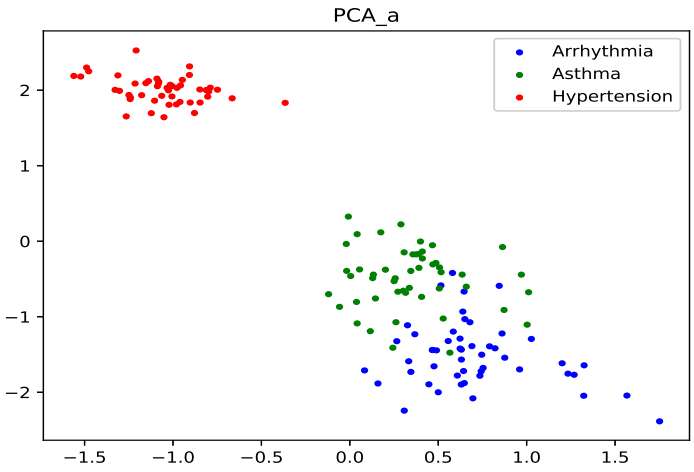
Singular Value Decomposition (SVD) shares the same idea with PCA, which is based on eigen values and eigen vector calculation.

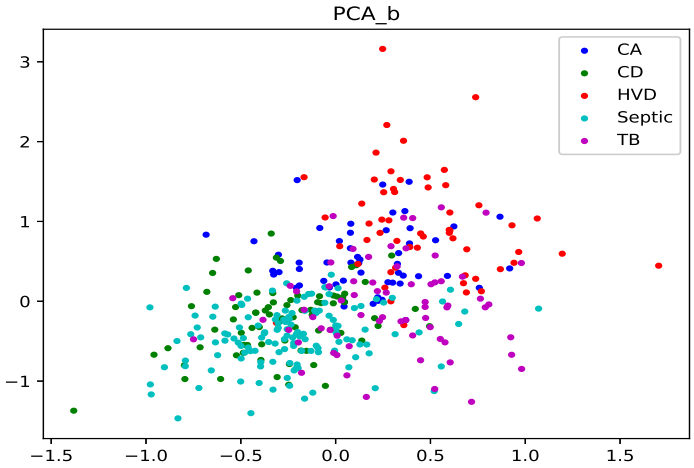
$$XX^\top = U\Sigma^2 U^T$$

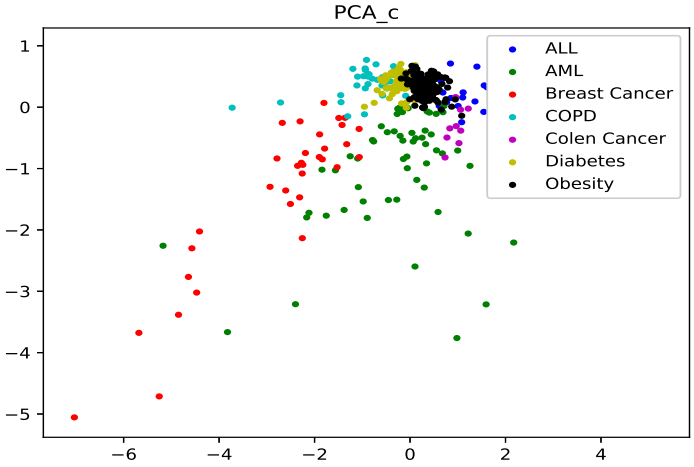
The equation shown above is the formula for calculation for covariance matrix. Σ actually represents the eigen value of the data. And we usually let the vectors from left hand side U be the eigen value. Due to the similar mechanism, the results obtained from both methods are quite similar.

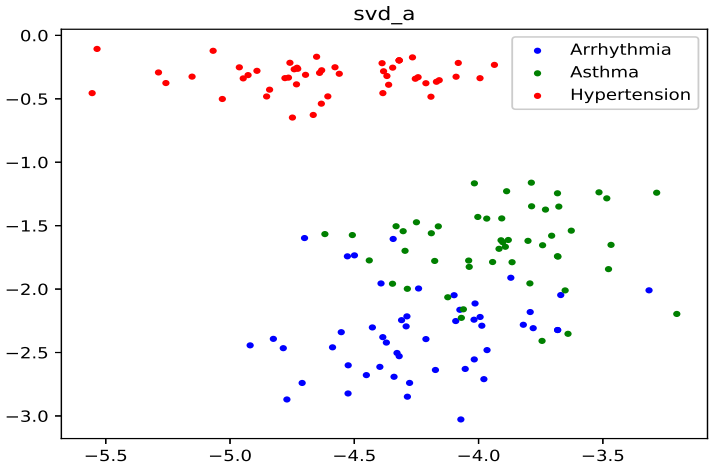
t-SNE converts the similarity relationship between sample points into probability: it is converted into probability based on Gaussian distribution in the original space(high-dimensional space);it is converted into probability based on t-distribution in the embedded space(two-dimensional space). This makes t-SNE not only pay attention to the local(SNE only pays attention to the similarity mapping between adjacent points and ignores the similarity mapping between the global, so that the visualized boundary is not obvious), and also pays attention to the global, so that the visualization effect is better(the clusters are not too concentrated, and the boundaries between clusters are obvious).But it takes much more time compared with PCA and more memories.

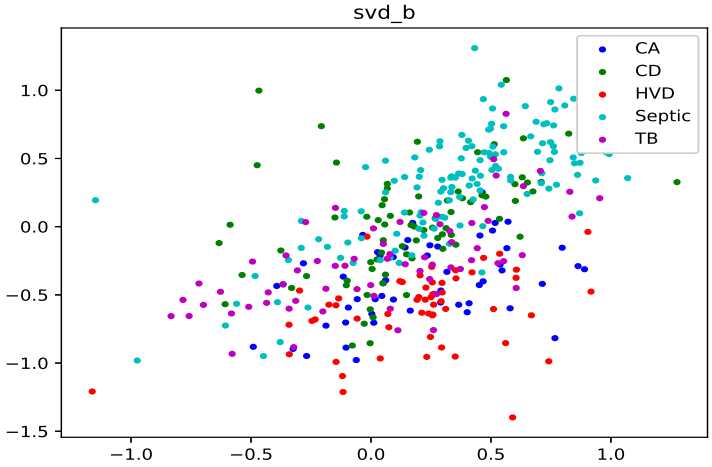
4. Appendix

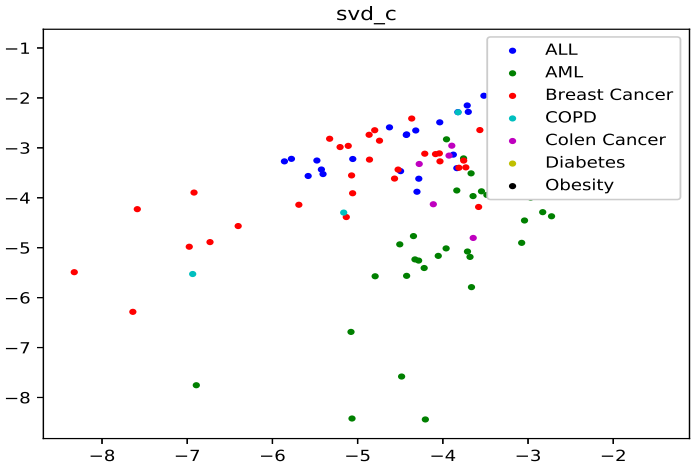


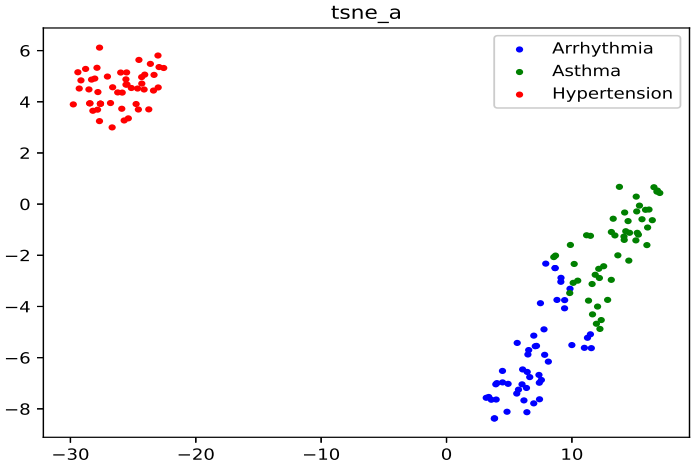

PCA_b
CA
CD
HVD
Septic
TB
3
2
1
0
−1
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5

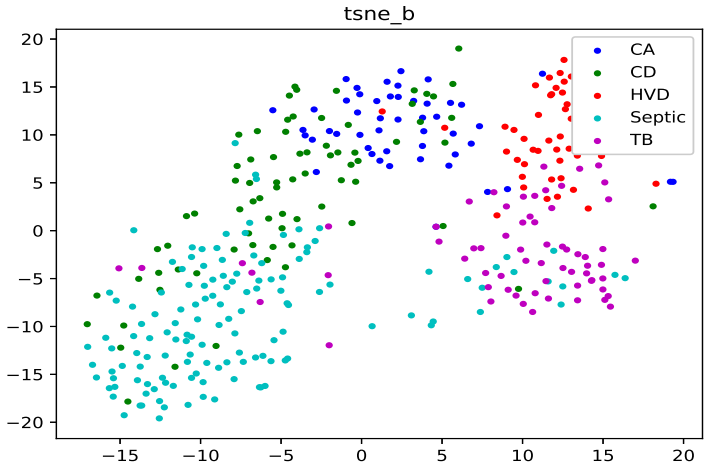

PCA_c
ALL
AML
Breast Cancer
COPD
Colen Cancer
Diabetes
Obesity
1
0
−1
−2
−3
−4
−5
−6
−4
−2
0
2
4

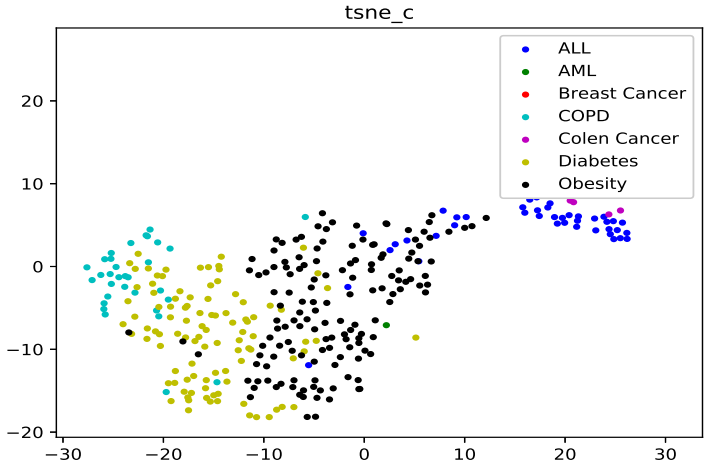

svd_a
Arrhythmia
Asthma
Hypertension
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
−5.5
−5.0
−4.5
−4.0
−3.5


svd_b
CA
CD
HVD
Septic
TB
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0


svd_c
ALL
AML
Breast Cancer
COPD
Colen Cancer
Diabetes
Obesity
−1
−2
−3
−4
−5
−6
−7
−8
−8
−7
−6
−5
−4
−3
−2


tsne_a
Arrhythmia
Asthma
Hypertension
6
4
2
0
−2
−4
−6
−8
−30
−20
−10
0
10


tsne_b
CA
CD
HVD
Septic
TB
20
15
10
5
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
15
20


tsne_c
ALL
AML
Breast Cancer
COPD
Colen Cancer
Diabetes
Obesity
20
10
0
−10
−20
−30
−20
−10
0
10
20
30