ACB-Assignment 5 and 6

Ankit Sharma - MT16121

November 12, 2017

1 Problem 1-RNA Sequencing[2]

1.1 Task 1 - Selection of RNA based study

Command for downloading SRA files:
wget - c ftp_path_to_the_sra_file
Command for obtaining the corresponding fastq files:
fastq-dump filename.sra

1.2 Task 2 - Perform quality check and generate sam files

Command for performing quality check:
fastqc filename.fastq
Command for downloading the index file:
curl -O ftp_path_to_the_index_file.fa.gz
Command for unzipping the downloaded file:
gunzip index_file_name.fa.gz
Command for building the index file:
bowtie2-build index_file_name.fa index_file_name
Command for generating SAM file:
bowtie2 -x index_file_name -U fastq_file_name -S output_file_name

1.3 Task 3 - Get raw data count using bam files and gtf file

Command for obtaining raw counts:
htseq-count output_file_name gtf_file_name

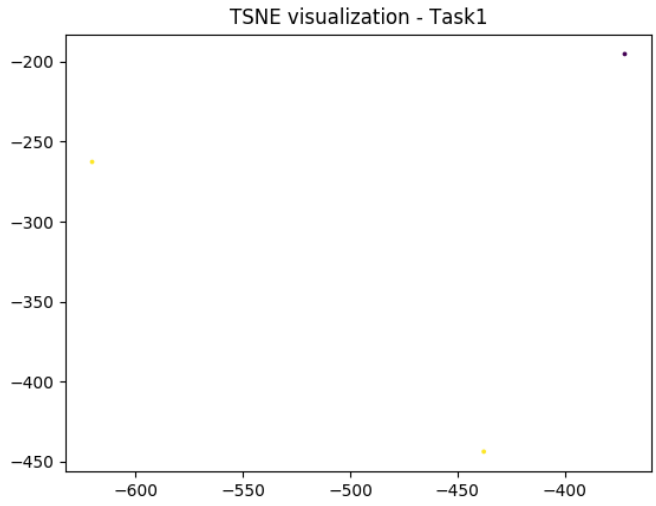
1.4 Task 4 - Perform normalization and clustering techniques

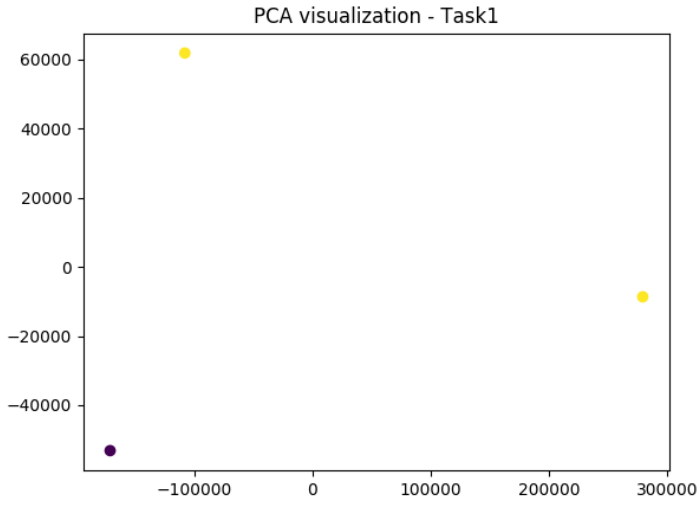
Geometric mean normalization was performed followed by KMeans clustering. Data was visualized after dimensionality reduction using PCA and was visualized using matplotlib. Details of the clustering algorithm as well as PCA is present in the next section of 'Clustering

on single cell data'. 100 % accuracy was obtained for KMeans clustering. The screen shot and plots are as follows:

```
C:\Python27\python.exe C:/Users/ankit/PycharmProjects/ACB_Assignment_5_6/MT16121_Ankit_Sharma_Task_1_2.py
Averaged accuracy for KMeans 100.0

Process finished with exit code 0
```





1.5 Task 5 - Significance of the biological data

The following SRA files for homosapiens were used:
1. SRR015164.sra
2. SRR015165.sra
3. SRR015166.sra
The study of which these samples are a part of is 'Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments'. Using these data files, methods were developed to predict alternate isoforms derived solely from exon expression levels in RNA-seq data. These three data points were clustered in two on the basis of literature. This dataset thereby aids alternative isoform analysis.

2 Problem 2-Clustering on single cell data[1]

This problem was coded using python 2.7 and the code is shared along with this report.

2.1 Task 1 - Normalization and log transform

The single cell study downloaded from recount database was 'Polymorphic cis- and trans-regulation of human gene expression'. This study dealt with 1,000 human genes with the motive to identify cis and trans acting polymorphic regulators. Two files were downloaded, 'cheung_count_table.txt' and 'cheung_phenodata.txt'. The first file consisted of a matrix in which rows represented genes and columns represented samples. The second file consisted of phenotype information for each and every sample which was used as the ground truth. The text files were read and a corresponding list of lists was created in python. The normalization technique used for this task was 'Geometric mean normalization'. To normalize data using this technique, each and every read count is divided by the geometric mean of the gene corresponding to that read count across all samples. This normalized matrix then underwent log transformation, in which log(1 + read_count) with base 2 was computed for every read count. This log normalized matrix was then written to 'log_norm.csv'.

2.2 Task 2 - Clustering

The three clustering techniques used were Spectral clustering, Agglomerative clustering and K Means clustering. The data was split into training and testing with the test data size as 30%. As the clustering depends upon the random split of data, therefore to ensure robust results, average accuracy was computed for 100 runs. Explanation about the clustering techniques used are as:

2.2.1 Spectral Clustering

Spectral clustering makes use of the eigenvalues of the Laplacian matrix to perform dimensionality reduction to a lesser number of dimensions. Once, the dimensions get reduced, the

points are clustered on the basis of the number of cluster centers that are provided as the input. Sklearn's "SpectralClustering" class was used to implement this kind of clustering with the number of clusters equal to 2. Predictions were obtained for the test input data and the predicted clusters were then compared with the ground truth for accuracy. An accuracy of 41.53 % was obtained using this technique. This technique was chosen as laplacian aims at capturing the neighbourhood of a data point which is an important aspect when we go for clustering.

2.2.2 Agglomerative Clustering

Agglomerative clustering is a special type of hierarchical clustering technique that makes use of the botton up approach. In this each data point begins in its own cluster and gets merged as we move up in the hierarchy. Sklearn's "AgglomerativeClustering" class was used for implementation with number of clusters equal to 2. The averaged out accuracy was 46.15%. Hierarchical clustering is best suited for genomic data that was the reason why I went for this technique.

2.2.3 KMeans clustering

This clustering technique aims to divide the data set into k clusters in which each data point belongs to a cluster with the nearest mean. The averaged out accuracy for KMeans clustering was 69.23%. Although this technique gave the highest accuracy but was the slowest of all. The algorithm took a lot of time to converge with respect to other two methods.

2.3 Task 3 - t-SNE and PCA

2.3.1 TSNE

TSNE stands for t-Distributed Stochastic Neighbor Embedding. This is a dimensionality reduction technique that is used for reducing very high number of dimensions to a lower number which can then be used for visualization. Sklearn's TSNE package was used for its implementation and then sklearn's matplotlib was used to visualize it. The plot is available in the following subsection. Data points are coloured on the basis of their clusters from the ground truth.

2.3.2 PCA

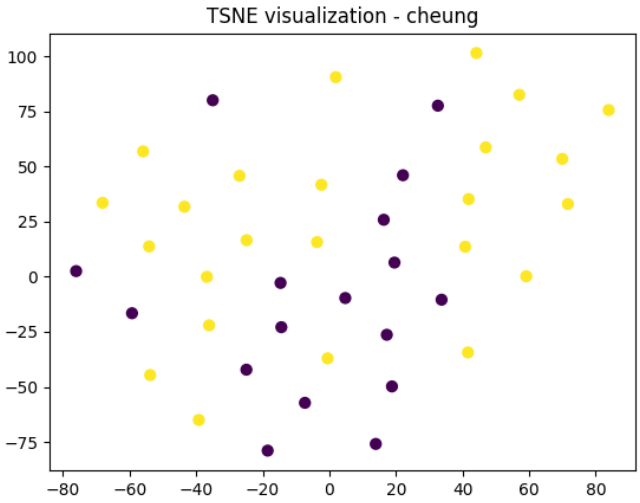
PCA stands for Principal Component Analysis. It is a statistical technique that uses orthogonal transformation to reduce dimensions. The first principal component has the highest variance, the second one has the second highest variance and so on. It is commonly used to visualize genetic distances among populations and that is the reason we have used it here. Sklearn's PCA class was used for the implementation followed by the used of matplotlib.

2.3.3 Accuracy and Plots

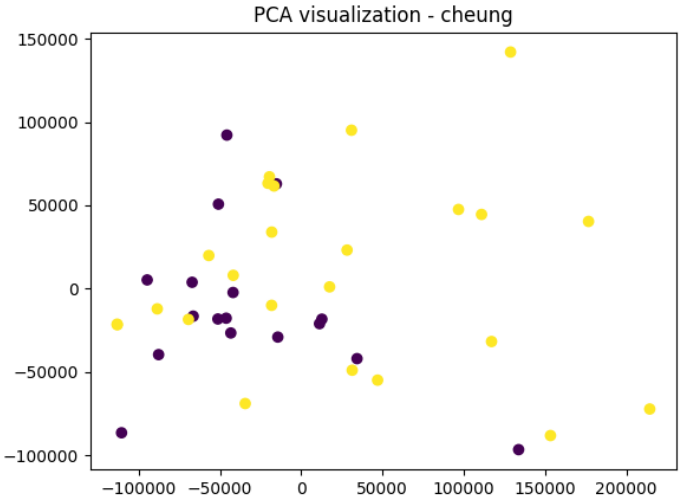
The screen shot for the accuracy is as follows:

```
Averaged accuracy for Spectral clustering 41.5384615385
Averaged accuracy for Agglomerative clustering 46.1538461538
Averaged accuracy for KMeans 69.2307692308
```

The plot for TSNE visualization is as follows:



The plot for PCA visualization is as follows:



2.4 Task 4 - Biological significance of the data

The data used in this study contains gene expression levels for various samples. There exists a direct correlation between gene expression levels and the phenotype exhibited but this has not been validated molecularly. Also, due to the complex nature of humans as biological systems, there exists a great amount of variability in the gene expression levels. This variation causes the human body to exhibit variations like differences in disease susceptibility etc. Thus, the biological data encountered in this study aims to bridge the gap in between genotype and phenotype.

3 Deliverables

All the code files, plots and scripts are checked in to BitBucket by the name of hw5_6_mt16121 and the repository is shared with ayalurarvind@gmail.com.

References

[1] Vivian G. Cheung, Renuka R. Nayak, Isabel Xiaorong Wang, Susannah Elwyn, Sarah M. Cousins, Michael Morley, and Richard S. Spielman. Polymorphic cis- and trans-regulation of human gene expression. *PLOS Biology*, 8(9):1–14, 09 2010.

[2] Hugues Richard, Marcel H. Schulz, Marc Sultan, Asja Nurnberger, Sabine Schrinner, Daniela Balzereit, Emilie Dagand, Axel Rasche, Hans Lehrach, Martin Vingron, Stefan A. Haas, and Marie-Laure Yaspo. Prediction of alternative isoforms from exon expression levels in rna-seq experiments. *Nucleic Acids Research*, 38(10):e112, 2010.