BIO310 Introduction to Bioinformatics
Homework 4 Spring 2021

May 24, 2021

Instructions:

- You are expected to program in Python. Upload the code together with your comments in `markdown` in the form of an `.ipynb` to SuCourse by the due date. Name your submission `BIO310-HW4code-YourName.ipynb` where you substitute in your first and last names into the file name in place of 'YourName'.
- For late submissions, please see the late submission policy in the syllabus.
- Please follow the submission instructions, not adhering the submission standards will lead to point deduction.

Introduction

In HW3, we used a dataset called `data.csv` that contains gene expression measurements (`gene_0, gene_1,..`) for different cancer patients (`sample_0, sample_1,..`) to cluster patients based on their gene-expression profile such that patients with the same cancer type were in the same cluster.

We also have another dataset called `labels.csv` that tells us which cancer type each patient has.

Using the same dataset, we now want to perform dimensionality reduction for data visualization. Since `data.csv` is high-dimensional, it is hard to plot in lower dimensions (for example in 2 dimensions along the x- and y-axis). In this assignment, you will use PCA and tSNE to reduce the dimensions in `data.csv` to 2. Then we plot this lower-dimension representation of `data.csv`, overlap it with the true (from `labels.csv`) and predicted (from k-means) clusters and generate a graphic to understand our findings better.

PCA + tSNE + Data Visualization [100 pts.]

1. **Generate predictions for clusters using k-means.** Use sklearn k-means to cluster samples based on gene expression levels. Set `k=5` and `init='k-means++'` for centroid initialization. Use `random state=1` parameter to produce the same results across different calls.

2. **Reduce dimensions of `data.csv`.**

   (a) Use sklearn PCA with `n_components=2`. What is the ratio of variance explained by the two components?

   (b) Use sklearn tSNE with `n_components=2`.

3. **Visualize k-means predictions together with true labels on the reduced dataset.**

   (a) Using the k-means algorithm from (1), get cluster number predictions for each patient.

   (b) Plot the results of PCA `fit_transform` in a scatterplot. Show the predicted and true labels in the same scatterplot.

   (c) Plot the results of tSNE `fit_transform` in a scatterplot, like you did for PCA.

   (d) Compare the two plots. What do you observe?