BIO310 Introduction to Bioinformatics
Homework 3 Spring 2021

May 6, 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-HW3code-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

We have provided you a dataset called `data.csv` that contains gene expression measurements (`gene_0, gene_1,...`) for different cancer patients (`sample_0, sample_1,...`). Using this dataset, we want to answer this question: which patients have similar cancer types? Is there a way to cluster the patients based on their gene-expression profile such that patients with the same cancer type are together?

We want to use `data.csv` and the k-means clustering algorithm to try and determine how many different cancers are included in the dataset. In other words, we will cluster the gene expression data `data.csv` using k-means, and try to predict the optimal number of cancers (clusters) k in the dataset, and which patients have the same type of cancer.

We also have another dataset called `labels.csv` that tells you which cancer type each patient has. We will use `labels.csv` to determine how good our clustering and predictions were.

Data [25 pts.]

Read in the data and answer the following questions related to it:

1. How many patients do we have data for?
2. How many genes are we measuring expression for?
3. Plot a histogram of the mean gene expression with 10 bins. What do you observe?
4. Which gene has the maximum mean expression? (Calculate mean expression of each gene, find the one with the max.)
5. How many unique cancers do we have in the dataset?

K-means Clustering [75 pts.]

K-means is a frequently used clustering algorithm which divides or partitions the data points into a pre-determined, "k" number of clusters [1]. In this part, we will pretend that we do not know the cancer labels for the samples and cluster patients.

1. Use sklearn k-means to cluster samples based on gene expression levels.
   Whenever you use the k-means, set `init='k-means++'` for centroid initialization. Use `random state=1` parameter to produce the same results across different calls.
   (a) Run the algorithm for different k values: `k= 2, 3, 5, 6, 7, 8, 9`. For each k run the algorithm 5 times. Report the best sum-squared-error obtained from multiple runs for each k in a table.
   (b) Using these, generate an elbow plot and pick the best value of k.
2. Evaluate the quality of the clustering:
   (a) Calculate average silhouette width for each `k`. Pick the best one.
   (b) Plot the silhouette width graph for the best value of `k`.
   (c) Does the best `k` (the optimal number of clusters) you have chosen match the true number of cancer types of the data?

Possible Workflow

If you need to, use the following to-do list to guide you along one possible workflow.

Most of the functions mentioned are from `python` libraries including `pandas`, `sci-kit learn`, which are very commonly used by professionals working with data. Familiarity with these will help you land internships and jobs in the field. So do your best with this assignment!

Every data science workflow comprises of some common steps: reading in, exploring, and pre-processing data followed by training and evaluating learning algorithm.

Below we go through some of these steps:

1. Read in data
   - ☐ The data contains 2 files: `data.csv` and `labels.csv`. Read in the data. One easy of doing this is to use the `pandas read_csv` function. Look at other arguments this function takes to make the input look cleaner/read different input types.
2. Look at data
   - ☐ Check out pandas functions available for viewing your data. Try them to get a feeling for what they do and what your data contains.
   - ☐ Often, some of the values might be missing. There are ways of finding these values and counting them.
   - ☐ Since the data is very large, just viewing a few rows or statistics might not be enough. Plotting the data in different ways is a good way of getting a better 'feeling' for the data. Here are some ideas of the kinds of plots you can create.
3. Pre-process data for machine learning algorithm
   - ☐ There are different ways of dealing with NA values. Here we'll just replace each NA value by 0. Related resource.
4. Train machine learning algorithm
   - ☐ Start working on clustering using K-means algorithm.

References

[1] Altuna Akalin. Computational genomics with r.