Computational Medicine: 02-518/02-718

Carnegie Mellon University

Homework 1

Version: 1.0; updated 9/9/2020

Due: October 4 by 11:59pm

Hand-in: A single PDF to Gradescope that contains the following items:

1. A cover page that lists your name and Andrew id
    - If you worked on a team, indicate who your teammate(s) is/are by their name(s) and Andrew id(s).
        - Each team should have no more than 3 people.
    - Note: Each team should only hand in one pdf. It does not matter who does the actual upload. We will make sure that the grades are entered appropriately.
2. A PDF export of the Jupyter notebook for question 1
3. A PDF export of the response to question 2.

You can combine all three PDFs into one using Adobe Acrobat, or similar tool.

Overview

In this assignment, you will perform cluster analysis to identify and characterize clinical phenotypes. It will give you the opportunity to apply concepts covered during lecture one [1], as well as during lectures 5-8 (weeks 3 & 4).

Notes:

- As is the case with real research, there are multiple ways to analyze the data and there may be multiple, equally valid answers. The goals of the assignment are as follows: 1) to give you the opportunity to perform cluster analysis on clinically relevant data; 2) to have you devise, apply, and justify a method for deciding how many clusters (i.e., phenotypes) exist within the data.
- You will not be graded on your choice of clustering method. You are free to use any method you wish.

Question 1 (90 points)

The first question involves clustering serologic data from COVID-19 patients. Download the file **HW1_data.zip** from the homework webpage. Extract the files from the archive. One of those files is named **HW1_Q1_data.csv**. That file contains a 232 by 61 matrix, plus a header row. The header row contains the names of serological markers, some demographic variables, some binary variables, and an outcome. These variables will discussed in-class. The remaining rows are the gene expression values from 232 COVID-19 patients.

To complete this question, use the provided jupyter notebook.

Part 1.1 (15 points)

Load and standardize **[2]** the data in columns 1-52. That is, subtract off the mean of each column and scale it to unit variance. Ex. if the raw value is x, and the mean and standard deviation of the column in which x is found are μ and σ, respectively, then the standardized value of x is $z = (x - \mu)/\sigma$. Note: standard scores are also called z-scores. The value z is simply the number of standard deviations x is above or below μ.

Select a clustering algorithm, distance measure, and a method for computing the quality of a clustering. You can use any clustering method you wish. Likewise, you can use any quality measure you want.

Apply the clustering algorithm to the **standardized data** in columns 1-52 (i.e., the real-valued variables) for $1 < k < 11$. Do not include the data in columns 53-61 when clustering.

How many clusters are there in the data? To justify your answer, create a plot where the x-axis corresponds to different values of k, and the y-axis corresponds to the quantity you chose for measuring the quality of each cluster.

Part 1.2 (15 points)

Perform a univariate analysis like the one performed in "*Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data*" J Allergy Clin Immunol. 2014, 133(5):1280-8 (see Lec. 5), and find 4 variables where there are statistically significant differences (i.e., $p < 0.05$ using pairwise *t*-tests) between the values in the clusters.

Create 4 box plots (one for each variable you selected), like those seen in Fig 1. from the paper by Wu *et al.*

Part 1.3 (15 points)

Summarize each of the clusters you identified in part 1.1 using the variables that *were not* used during clustering (i.e., columns 53-61). **Create a table** where the rows correspond to the variables in columns 53-61, and the columns correspond to the *k* clusters you identified. For each cell in the table, put summary statistics for that (variable, cluster) pair. For example, the number of men vs the number of women in each cluster.

Are any of the clusters significantly enriched for some particular value? For example, are there clusters that consist primarily of women? Can you show that the enrichment is statistically significant?

Part 1.4 (15 points)

There are 52 serological measurements per patient. Each measurement requires an experiment, which can be expensive. In this question, we will attempt to reduce the number of variables by following one of the strategies used in the paper "*Unsupervised phenotyping of Severe Asthma Research Program participants using expanded lung data*" J Allergy Clin Immunol. 2014, 133(5):1280-8 (see Lec. 5).

Cluster the numeric variables (columns 1-52). Select a clustering algorithm, distance measure, and a method for computing the quality of a clustering. The authors of the papers used hierarchical clustering, but you are free to use any clustering method you wish. Likewise, you can use any quality measure you want.

How many clusters are there among the variables? To justify your answer, create a plot where the *x*-axis corresponds to different values of k

, and the y-axis corresponds to the quantity you chose for measuring the quality of each cluster.

Part 1.5 (15 points)

Select a representative variable from each cluster you identified in part 1.4, and then create a low-dimensional version of the data set using those variables. Re-cluster the data using that reduced representation using the same choices you made for part 1.1.

How many clusters are there in the reduced data? To justify your answer, create a plot where the x-axis corresponds to different values of k, and the y-axis corresponds to the quantity you chose for measuring the quality of each cluster.

Part 1.6 (15 points)

Summarize each of the clusters you identified in part 1.5 using the variables that *were not* used during clustering (i.e., columns 53-61). **Create a table** where the rows correspond to the variables in columns 53-61, and the columns correspond to the *k* clusters you identified. For each cell in the table, put summary statistics for that (variable, cluster) pair. For example, the number of men vs the number of women in each cluster.

Are any of the clusters significantly enriched for some particular value? For example, are there clusters that consist primarily of women? Can you show that the enrichment is statistically significant?

Question 2 (10 points)

In this question, you will create a Concept Map describing the relationships between terms and concepts relevant to the module on phenotyping (lectures 5-8). Here is an example of a Concept Map.

The focus questions for this Concept Map are:

- "*What are clinical phenotypes?*", and
- "*How do we identify them from clinical data?*"

Normally, the process of creating a Concept Map begins with the creation of a list of concepts. Since this is the first assignment, and you may not have experience with Concept Maps, we will get you started by giving you some concepts. **Feel free to add concepts to this list.**

Concepts:

- Agglomerative clustering
- Behavioral Factors
- Centroid-based
- Clinically Relevant
- Cluster Analysis
- Cluster evaluation measures
- Density Based
- Disease
- Distance Function
- Divisive clustering
- Environmental Factors
- External
- Genetic Factors
- Hierarchical clustering

- Internal
- K means clustering
- Medical Record
- Phenotypes
- Precision Medicine
- Subtypes
- Signs
- Spectral
- Symptoms

You **do not** need to incorporate all of the concepts in this list into your concept map, but try to include at least 10.

For this question, you will be graded on whether you demonstrate an understanding of the relationships between the concepts you include in your map (by adding connections and labels on those connections). Points will be deducted if you make “incorrect” connections. In future assignments, you will also be graded on the list of concepts you generate.

Use Powerpoint (or some similar tool) to create the concept map. Save the map as an image or pdf, and include it with your handin for HW1.

---

[1] Recall the example during lecture one from "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications" Sorlie, et al PNAS 2001 11;98(19):10869-74

[2] It is often a good idea to standardize raw data prior to clustering (or other kinds of analysis), especially if the raw data values are on very different numeric scales and/or have very different variances. In the context of clustering, standardizing addresses some of the issues associated with computing distances over arbitrary values.