Mini Project 2

EN 601.751, Fall 2017

Prof. Alexis Battle

DUE DATE: 11/27/2017

1. Writing methods text:

- For a paper we have read in class where you thought the methods were unclear (any paper up through 11/3), rewrite a key piece (1-2 paragraphs of the stats/ML part) of the methods section to improve clarity, drawing on your understanding of the method from the class discussion. If there is something you still don't fully understand, feel free to simply mark it, I'm primarily looking at this as a writing exercise, not to penalize if you get a detail wrong.

2. Reproducibility:

- Pick any paper we have read in class, from the list at https://sites.google.com/site/en600641/readings, OR one of your own interest but relevant to class, and spend no more than 4 hours trying to reproduce one of the results. A simple/initial part of the results is fine.
- This can also serve as a starting place or preliminary exploration for your final project (where one project option will be to replicate part of a paper and add some small extension of your own). Feel free to pick a paper you'd like to extend.
- Guidelines:
    - **Be sure you are selecting a paper with public data - it doesn't count if you simply can't access the data at all because it's under access protection.**
    - Synthetic data (as described by the author) IS allowed for this exercise.
    - Known public data include: GTEx, CCLE, TCGA, GEO, GEUVADIS etc, so papers using the public parts of these datasets will tend to be easier.
    - **You will not be penalized for difficulty in replication, e.g., if the data is complex to process and filter, if the description in the paper is too vague, or if the analysis is complex to replicate – the exercise is simply to see how far you can get in four hours.**

2.1. Deliverables:

- Document the process you went through step by step – describe what you did, any challenges you faced etc.
- Send any code/figures that result from your efforts, including processing scripts

3. Final project proposal

Read the final project specifications, and write a short paragraph describing your chosen project. By the time of the proposal you should have identified the data you will work with verified that you can appropriately download/access and understand the data, and produced a simple analysis (described below). Sanity-checking at this early stage will save you the pain of realizing the data will not serve your purposes too close to the final deadline.

Preliminary analysis: provide 2 simple figures summarizing important QC or basic relationships in the data you plan to use. Examples could be initial clustering of data, PCA plots, histograms summarizing distribution of important quantities etc. Essentially you want to make sure you can process your data, and that the signal you are interested in does exist in this data. If you used the same data in Exercise 2, one figure may be reused between Exercises 2 and 3.