Topics in Biomedical Data Science: Large-scale inference (BIODS215-2018) Problem set 2

Due date: 2018/2/22, Release date: 2018/2/6

Yosuke Tanigawa, Julia Salzman, James Zou, and Manuel Rivas

- Please write the answer to the problem set in one pdf document that includes all the codes and results from computational experiments.
    - Jupyter Notebook [1] and R Markdown [2] have functionalities to export a document into a pdf file.
- Please submit your answer through gradescope

1. Data visualization for high dimensional data (20 pts)

Please briefly describe the following topics.

a. Why high dimensional data is hard to approach (compared to one- or two-dimensional data)? Do you think having domain knowledge can help us understand data in high dimension? Why or why not?

b. What is the difference between principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE)? What are the advantages and disadvantages of the two methods. In what situations one method is preferable to the other?

Please apply hierarchical clustering with different distance metrics for expression data.

c. Downloaded a gene expression dataset (microarray data or RNA-seq data) of your choice from GEO (Gene expression omnibus) database.
Please apply hierarchical clustering algorithm (you may use libraries/packages of your choice. You don't need to implement the clustering algorithm) and visualize the results with a heatmap and dendrograms.
Try different distance (Euclidean, Manhattan, Chebyshev norm, etc.) and different cluster distance (single linkage, complete linkage, and UPGMA) and report how the results change. How do you interpret the results? What is the limitation of the analysis?

Note: Typically, microarray data is in CEL format and R Bioconductor package is useful to process this data. Please refer to the reference & resource section.

2. **Bayesian multi-level modeling using real-world data** (60 pts)

   This is about multi-level meta-analysis using summary level data from Global Biobank Engine (https://biobankengine.stanford.edu) and IBD exomes browser (http://ibd.broadinstitute.org). The goal of this problem is to teach you how to perform summary level regression and summarize parameters of interest.

   a. Install the `brms' package (https://github.com/paul-buerkner/brms). If having difficulty installing the brms package with its dependencies please contact the TA, Yosuke Tanigawa.
   b. Read in data on estimated effects for rare variants in *NOD2* and estimated effect on crohn's disease and ulcerative colitis.
   c. Summarize parameters like correlation of genetic effects across crohn's disease and ulcerative colitis.
   d. Summarize posterior effect size estimates for individual level genetic variants.
   e. Summarize parameters for `missense` and `lof` and `silent` variants.
   f. Provide a summary of the posterior distribution for a `missense` allele in *NOD2* with estimated odds ratio of 1.5 and standard error of 0.05
   g. Provide a summary of the posterior distribution for a `missense` allele in *NOD1* with estimated odds ratio of 1.5 and standard error of 0.05.
   h. Plot your posterior predictive distributions.

3. **Prediction with linear models** (10 pts)

   a. Replicate the analysis in lecture material 11 using Bayesian linear model
   b. Provide prediction and uncertainty estimate of the same prediction using classical linear model
   c. Compare the two results and describe what can you learn from the side-by-side comparison of the two regression methods.

4. **Deep Learning in genomics** (20 pts)

   This list https://github.com/hussius/deeplearning-biology#genomics is a collection of deep learning research for genomics. From this list, please select one paper which is a good application of deep learning and one paper where you believe deep learning is not appropriately used. Explain and support your rationale with discussion of specific results from each paper (one or two paragraphs per paper).

Reference & Resources

1. Jupyter notebook https://http://jupyter.org
2. R Markdown http://rmarkdown.rstudio.com
3. Gradescope https://gradescope.com , Entry Code:MWYKE4
4. Gene Expression Omnibus (GEO) datasets https://www.ncbi.nlm.nih.gov/gds
5. R Bioconductor https://www.bioconductor.org/
6. Example workflow for the microarray analysis https://goo.gl/t2sisA
7. Global Biobank Engine https://biobankengine.stanford.edu
8. IBD exomes browser http://ibd.broadinstitute.org
9. R brms package https://github.com/paul-buerkner/brms
10. Deep Learning in genomics https://github.com/hussius/deeplearning-biology#genomics