Topics in Biomedical Data Science: Large-scale inference (BIODS215-2020) Class project

Proposal due date: 2020/1/23, Release date: 2020/1/9

Yosuke Tanigawa, James Zou, and Manuel Rivas

We think the class project is a great opportunity for you to use some of the methods you will learn in the class. To allocate sufficient time to work on the project, we would like to have a brief **project proposal** by the third week of the quarter, **1/23/2020**.

Please briefly describe your proposal for the class project. Specifically, **your proposal must include the followings**:

1. Research question and/or objective(s)
2. Dataset
3. Types of statistical and/or deep learning models in your mind (if possible)

It might be hard for students to narrow down a specific model because we haven't covered most of the materials yet. If you're lost, we would recommend you to come to the office hours or reach out to the teaching team. We are happy to help you.

You can work on your original ideas for the course project. In case you don't have a specific project in your mind, the teaching team will provide you some useful datasets and project ideas. Each instructor will present their project proposal and the relevant datasets in their first lectures. The TA (Yosuke) is available to help you to narrow down your class project.

You can work as a group (**2 people maximum per group**). Please clarify the contribution of each group member in the final project write-up.

If you are working in a lab, you may use some of the results from your research as the class project, given that PI(s) are aware and supportive about it. If you were to take this route, we will ask you to clarify the relevance of the project to the materials covered in the class.

Some project ideas and available datasets

Project proposal by Dr. Manuel Rivas

1. Bayesian models for functional enrichment analysis
2. Statistical models for disease prediction
3. Bayesian models for rare variant association studies
4. Predictive Gaussian Process models for wearable data
5. Building predictive models from imaging data and tying genetics

Available datasets through teaching team and/or publicly-available datasets

While most of the datasets listed here are publicly-available, some data needs to have restricted access. If you're interested in the datasets with the restricted access, please contact the TA (Yosuke).

1. Epigenomics data from fetal tissues, including matched ChIP and RNA-seq data
    a. Raw and processed data from Roadmap epigenomics data that contains ChIP-seq, RNA-seq, and variants during fetal development
    b. Processed dataset for the GARFIELD pipeline (GARFIELD doi:https://doi.org/10.1101/085738).
    This is essentially a binary matrix of size (# variants) x (# epigenomic features)
2. TCGA (the cancer genome atlas)
3. Datasets from the UK Biobank project.
    a. GWAS summary statistic data for various diseases
    b. Imaging datasets
4. The latest dataset from the iPOP project (http://snyderome.stanford.edu/). This dataset includes wearable data.
    a. The wearable dataset can be found here: http://ipop-data.stanford.edu/wearable_data/Stanford_Wearables_data.tar
5. Transcription factor binding sites
    a. The ChIP-Atlas (https://chip-atlas.org/) has pre-processed ChIP-seq tracks.
    b. The GARFIELD data has some binding site information (5-6 TFs).
    c. One can process ChIP-seq data from the ENCODE/Roadmap project with peak calling software like MACS2 to identify binding peaks of transcription factors.