Project for MDGE 612

**Predicting phenotype by genotypes: a comparison of multiple models**

Please download the genotype file (genotype.csv) and the phenotype file (FT10.txt) from the D2L. Try to predict the phenotype based on genotypes. You may use any three methods learnt in the course to carry out the predictions and compare the powers of them. It is suggested to use three linear models: (1) polygenetic linear model; (2) L1 regularization (LASSO); and (3) L2 regularization (Ridge Regression). Please note that, for any of these methods, you need to carry out genome-wide association mapping first to narrow down candidate SNPs. In particular, for linear model, the number of SNPs has to be significantly lower than the sample size.

Please (1) identify your research goal, (2) design your pipeline of the study, (3) implement the pipeline, (4) conduct the analysis, and (5) interpret the results. The gene models (GFF file in D2L) and online information regarding the annotation of the genes may provide helpful information regarding how to select candidates. However, it is totally fine if you don’t use the biological annotation and deem this question as a pure “black-box” machine-learning problem.

Finally, write a report on the outcome of your research project. Please provide sufficient details to ensure that your outcome is replicable by others who may want to follow your work. The report should be professionally written towards the line of a scientific manuscript. It should include the components of a brief paper, such as introduction, methods, results, and discussion.

Grading rubric of the final report:

* (20%) Provide the rationale for the proposed study design.
  + Background in the field.
  + What is the goal of this work?
* (20%) The criteria for evaluating competing models.
  + Clearly state the criteria for evaluating the performance of the competing models
  + Multiple criteria might be used (e.g., how many SNPs should be used to start the comparison). If they lead to inconsistent conclusions, explain why this is the case.
* (30%) Methods and pipelines are clearly presented so that the readers can replicate the whole work based on the same data and your report. This includes:
  + A general illustration of the pipeline should be concise and to the point
  + Detailed codes are attached, clearly linked with the general illustration.
  + If existing tools (e.g., weka or an R package) are used, please specify the parameters used.
* (30%) The main conclusions are clearly stated and well interpreted. This includes:
  + Figures are well polished and figure captions are informative.
  + Statistical justifications are sound.
  + The texts justifying the findings should integrate figures and statistics.