Network Tour of Data Science project overview

Genetically determined susceptibility to disease

Valérian Rey, Rayane Laraki, Maxence Jouve, Artur Szałata

Lausanne, 4.12.2019

1 Objective

We aim to establish a link between genomes and immunity to certain diseases in mammals, in particular: influenza virus and malaria. We will attempt at predicting the response to the viruse and the malarial parasite given partial information about genotype and certain genes' expression in mice. Such knowledge, after extensive research, could be extrapolated to humans and those more susceptible could be advised to change lifestyle and have more frequent screenings to reduce disease risk.

2 Data description

We will use a dataset of genes, protein expressions and phenotypes of mice of BXD strains. It is a subset of the open dataset available at the genenetwork website. The data we will use consists of:

- Phenotypic data regarding the immune system. In the dataset "Phenotype.txt" marked with "Immune" category.
- Genotypes of those mice for which we have the phenotype of interest. We will extract the data from the "genotype_BXD.txt" We will narrow down the genes to the subset that is present in most of the mice under consideration.
- Protein expression data: subset of those related to the disease, basing the choice on domain knowledge and statistical correlation.

3 Approach

1. First we will establish a baseline for the graph based approach, that is try to predict the immune response directly from the genotype and the genes' expression.
2. Build two graphs: each one having mice as nodes and different phenotypic information as the predicted values, but differing in the distance metric determining the edges and the signal per node. In the first graph the difference in genotype is used as the distance metric between the nodes and the gene expression is the feature of the node, while in the second graph the features and distance metric are reversed.
3. Using those networks we will apply techniques for label propagation to fill in the missing values in genes and gene expression. We propagate target values (phenotype).
4. We try to predict the immune response of the mice using the baseline approach from point 1. on the data with propagated values and compare the results with the ones without the propagated labels. Improvement in the prediction would mean that we correctly propagated the phenotype.

4 Possible issues

- There is no established method for computing the similarity between genomes or genes' expression. This might render our network meaningless.
- The dataset has missing fields which might make the analysis and results not relevant.
- We have very limited number of data points. It is not desirable for machine learning techniques.