Final Project Proposal

The project is inspired by a real data analysis project, but for this assignment, I plan to compare the different feature selection methods/pipelines and see how the machine learning methods perform on the selected features obtained from such pipelines.

Dataset description

The data comes from PPG-DS study (Columbia University) in which patients with Down Syndrome were followed longitudinally over the duration of 9 visits (each visit 1-1.5 years apart).

Untargeted plasma metabolomic data were obtained from 297 of the patients in the first few visits. There are measurements on about 3000 metabolites' expression levels for each subject.

The outcome is incident Alzheimer Disease (AD), which is obtained from clinical diagnoses (gold standard). Among 297 patients with available metabolomic data, 267 have clinical diagnoses.

The goal of the whole project is to build a classifier that can use a reasonable set of metabolites to classify subjects with incident AD vs. no AD. Although there might be a weak relationship between baseline metabolites and future AD onset, it could be worthwhile to pinpoint a subset of potential metabolites for clinical diagnosis/prognosis and explore their bodily functions. The original project also aims to construct a risk score based on the selected metabolites and use it to model time-to-event. However, for this project, I plan to look at the feature selection pipelines. Specifically, I want to see if it is worthwhile to consider a pipeline starting with univariate filtering (e.g., t-tests), followed by multivariate filtering (PLS, LASSO, …), or just use one of the two, and which pipelines/methods work best given this metabolomic data. Prediction accuracy and time will both be taken into account when evaluating each pipeline. If time allows, more effort will be given to making sense of the results (using hierarchical clustering and/or graphical models on the selected set of metabolites).

The feature selection methods I consider using in the project are:

1) Empirical Bayes t-test in limma with FDR-adjusted pvalues <0.05
2) Linear model adjusted for surrogate variables in limma with FDR-adjusted p-values < 0.05
3) AUC filter using filterVarImp function in caret
4) PLS
5) LASSO/Elastic Net feature selection

These can be used as filtering methods (both univariate – 1,2,3 and multivariate – 4,5). Due to time limit, wrapper methods such as recursive feature elimination will not be considered.

Potential pipelines are as follows:

a. 1 or 2 or 3 -> 4 or 5 -> model
b. 1 or 2 or 3 -> model
c. 4 or 5 -> model

The models I plan to use on the selected features are SVM radial kernel, RF, XGBoost.

Some other specifications:

Since there is imbalance in the outcome (25% of subjects have AD), I will use ROC AUC as the metric and set the argument sampling = "up" to up-sample the minority class. Unit-variance scaling will be used on the metabolites. Models will be trained using 10-fold CV, repeated 5 times.