Project Title: Modeling Complex Genomic Associations

Project Author: Ryan Urbanowicz – Cedars Sinai Medical Center

Short Description: This project deals with binary classification tasks (case/control) in a variety of simulated single nucleotide polymorphism (SNP) datasets. Each dataset has a different form of underlying complex association (e.g. multivariate additive, epistatic, or genetic heterogeneity). This project will introduce participants to topics such as basic data preparation, feature selection methods, machine learning modeling methods, automated machine learning, and model interpretation.

Suggested Tags: binary, classification, machine learning, feature selection, automl

Long Description:

This this project we have provided a small variety of simulated genomic single nucleotide polymorphism (SNP) datasets that include a binary class outcome. The intended goal of this project is for students to apply feature importance, machine learning modeling algorithms, or any other statistics, machine learning, data science, or artificial intelligence methods they think may be relevant to train and evaluate predictive model(s) that achieve the best prediction performance possible as well as identify strategies to correctly distinguish between predictive and non-predictive features in the respective datasets.

Through examining the different datasets, students are encouraged to consider the unique challenges presented by different unique patterns of association in data (i.e. additivity, epistasis, and genetic heterogeneity). What modeling algorithm can detect these associations, and what processing steps can be taken on these datasets (with a different number of total features) to improve the ability of modeling to detect and interpret these effects.

Datasets:

Datasets were simulated using the GAMETES software package. In each dataset, most features are non-predictive (i.e. randomly simulated based on a randomly chosen minor allele frequency between 0.01 and 0.5. The remaining (predictive) features have been simulated to have a unique association that is predictive of a binary outcome in the dataset column named 'Class'. This outcome has been encode as 0 or 1 representing control subjects or case subjects, respectively. Non-predictive features start with the letter 'N', and predictive features start with the letter 'M'.

8 datasets have been simulated in total, each with 1000 samples/instances, and a degree of 'noise' which should make it impossible to predict any hold-out testing data with 100% accuracy. 4 'small' datasets have been simulated with a total of 100 features, and 4 corresponding 'large' datasets have a total of 100,000 features. For each, the 4 datasets include a different underlying 'pattern of association', i.e. the association between the predictive features and outcome is of a different nature.

4-wayAdditive: 4 predictive features that have been additively combined to predict outcome, such that all predictive features have univariate associations with outcome.

2-wayEpi: 2 predictive features that have a 'pure' epistatic interaction predicting outcome. Neither feature will have a univariate association with outcome.

2Additive_2-wayEpi: 4 predictive features total, representing two separately simulated 2-feature epistatic interactions that have been additively combined to predict outcome. Features in this dataset can potentially have both epistatic and univariate associations with outcome.

4-wayHeterogeneous: 4 predictive features that are each predictive only within a respective ¼ of all instances in the data (i.e. they are heterogeneous associated with outcome). Each feature has a weaker univariate association with outcome.

References and Suggested Reading:

- Urbanowicz, R.J., Kiralis, J., Sinnott-Armstrong, N.A., Heberling, T., Fisher, J.M. and Moore, J.H., 2012. GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. BioData mining, 5, pp.1-14.
- Urbanowicz, R.J., Kiralis, J., Fisher, J.M. and Moore, J.H., 2012. Predicting the difficulty of pure, strict, epistatic models: metrics for simulated model selection. BioData mining, 5(1), pp.1-13.
- Woodward, A.A., Urbanowicz, R.J., Naj, A.C. and Moore, J.H., 2022. Genetic heterogeneity: Challenges, impacts, and methods through an associative lens. *Genetic Epidemiology*, *46*(8), pp.555-571.
- Urbanowicz, R.J., Meeker, M., La Cava, W., Olson, R.S. and Moore, J.H., 2018. Relief-based feature selection: Introduction and review. *Journal of biomedical informatics*, *85*, pp.189-203.
- Urbanowicz, R.J., Olson, R.S., Schmitt, P., Meeker, M. and Moore, J.H., 2018. Benchmarking relief-based feature selection methods for bioinformatics data mining. *Journal of biomedical informatics*, *85*, pp.168-188.
- Urbanowicz, R., Zhang, R., Cui, Y. and Suri, P., 2023. STREAMLINE: A Simple, Transparent, End-To-End Automated Machine Learning Pipeline Facilitating Data Analysis and Algorithm Comparison. In Genetic Programming Theory and Practice XIX (pp. 201-231). Singapore: Springer Nature Singapore.