**MACHINE LEARNING IMPUTATION ANALYSIS WITH HIGH DIMENSIONAL BIOLOGICAL DATA: METHYLATION ARRAY**

A large portion of time spent working on a machine learning project is most likely spent on data preparation. This can be in the form of exploratory data analysis, which may lead to initial feature and sample filtering. Some datasets can be incomplete and riddled with missing data. This missing data poses a problem to the rest of the analysis, including most machine learning algorithms/applications. Many solutions exist to the issue of missing data, such as just removing offending samples, replacing missing values with either the mean, median, or mode of that feature, or even neighbor-based approaches like k-Nearest-Neighbors (kNN). Unfortunately, high dimensional data seems to benefit the least from these solutions. Biological data tends to be feature-rich, providing this unique issue of dealing with missing data in high dimensional datasets. This project will be focused on achieving high accuracy with imputation with methylation array data found on the Gene Expression Omnibus (GEO) using multiple machine learning methods found in literature. This includes HoloClean AimNet attention-based learning1, 2, Miss-Forest3, and decision trees and fuzzy clustering with iterative learning4.

High dimensional data suffers from the curse of dimensionality6, in which approaches to imputation with low dimensional data begin to show poor results as dimensions increase. Sometimes biological data such as methylation array data tends to be high-dimensional and have many missing values. This provides a unique issue of dealing with poor imputation, sample/feature removal, and a combination of the two. To try and help reveal a better imputation strategy for missing data in high-dimensional datasets, this project will compare multiple machine learning methods for imputation of missing data.

To begin, I will use the Gene Expression Omnibus to find a large set (500+ samples) of Illumina EPIC methylation array5 data (~800,000 features/probes). I will then perform some exploratory data analysis (EDA) and quality control (QC), such as finding redundant probes, finding correlations, covariances, and possibly poor-quality samples/probes. After EDA and QC, I will perform baseline methods for imputation, such as mean, median, and kNN. I will then try to implement HoloClean AimNet attention-based learning1, 2, Miss-Forest3, and decision trees and fuzzy clustering with iterative learning4. Although I am not sure about being able to implement all of these algorithms, I will still be able to compare any new imputation methods I implement to the previously mentioned forms of simple imputation.

Initially, I will take about a week to find the methylation array data on the Gene Expression Omnibus. Then, it will take about a week to do some exploratory data analysis and quality control, as well as perform the baseline methods for imputation. Then, I will spend most of my time, 2 weeks trying to understand and get each ML Imputation method working, starting with HoloClean AimNet. The final week/half week I will spend putting together the report, along with preparing any last notebooks for easy running/understanding. By this time, I will have succeeded with at least one of the ML models above for imputation, and will be able to compare that method directly to kNN, mean/median/mode imputation.

Citations

1. <https://cs.uwaterloo.ca/~ilyas/papers/WuMLSys2020.pdf>
2. <http://www.vldb.org/pvldb/vol10/p1190-rekatsinas.pdf>
3. <https://academic.oup.com/bioinformatics/article/28/1/112/219101>
4. <https://sci-hubtw.hkvisa.net/10.1007/s10115-019-01427-1>
5. <https://www.illumina.com/products/by-type/microarray-kits/infinium-methylation-epic.html>
6. <https://link.springer.com/referenceworkentry/10.1007%2F978-1-4899-7687-1_192>