Feature Selection + Classification

Domain and Data

MADELON is an artificial dataset, which was created for a feature selection challenge. The difficulty is that this dataset has 500 features and they are highly non-linear.

Instances: 2000
Features: 500

Problem Statement

Our firm is bidding on a big project that will involve working with thousands or possibly tens of thousands of features. Conventional feature selection techniques are impossible to use.

Solution Statement

I propose that a way to win the contract is to demonstrate a capacity to identify relevant features using machine learning. We can build pipelines to chain the steps of data processing then search through all parameters for the best model.

Metric

We will use the mean accuracy score to compare the models and selected features.

Benchmark

Our benchmark accuracy is 0.85 using SelectKBest and KNeighborsClassifier in a pipeline.

Phase I - "Benchmarking" - we build the necessary wrapper functions to execute our project. Once that was completed, we used our algorithm to benchmark the performance using LogisticRegression at default levels. Our base accuracy was 0.534.

Phase II - "Identify Salient Features Using $\ell 1$-penalty" – we narrowed down our features by adding L1 penalty. The L1 penalty will eliminated irrelevant features by penalizing their coefficients. 10 features were extracted to be relevant out of the 500.

Phase III – "Build Model" – we automated and optimized our models by running them through a pipeline then going through parameters to find the best ones. The LogisticRegression pipeline accuracy score was 0.61, which didn't improve any more than the last phase. However, it was greatly useful to quickly identify salient features. 2 Algorithms SelectKBest and KNeighborsClassifier gave the best accuracy score of 0.852. We can use this as our 2nd benchmark and further examine optimization.