Automatic Feature Selection for Large Sets of Features

This possibility of automatic feature selection is explored as a proof of concept. The potential client has a dataset that has a very large number of features. Without using their data, an alternate dataset can show the viability of automatic feature selection.

A canonical dataset with a large number of features, mostly irrelevant features, is the 'madelon' dataset. It has a binary target feature, 5 salient features, 15 more feature based on the 5 salient features and 480 irrelevant features. If the relevant features can be automatically and accurately selected. The process could be generalized to other datasets. The binary target feature requires us to use logistic or classification regression.

The metric for evaluating the models is accuracy. Other metrics would not make sense for use with a discontinuous target, like this dataset uses.

The most basic benchmark for this dataset is 50% accuracy because the dataset is evenly split on target feature. Better benchmarks for accepting a good model would be accuracy greater than 75% with less than 20 features.

RESULTS:

The first step was to conduct a logistic regression to set a baseline for performance of the model on this dataset. The full dataset was evaluated and baseline accuracy for the model was set at 52%.

The dataset is sufficiently noisy to require more complex models for making a prediction.

The second step was to add a 'l1' penalty to regression. Again, the dataset was too noisy to effectively be evaluated. The accuracy was about 50% and only 25 features were eliminated using this technique. This reduction in features is not useful for any comparison.

The third step used K best feature selection, K Nearest Neighbors and a grid search to find a better model of feature selection. Using the K Best transformer reduces the feature set to help find the best candidates for salient features. Combining the K Best transformation and a logistic regression improved the accuracy to about 60%. Since this is below the benchmark, other models need to be explored.

The k nearest neighbors (KNN) model showed sharp jump in accuracy without altering the default parameters. Using a grid search, the accuracy of the KNN model achieved 88% accuracy on the test set. The model used 14 features to create this model.

Moving forward with this data set, a layer of logistic regression on top of the KNN model could help identify the 5 salient features. Since some the features identified by the KNN model are likely to be combinations of the salient features, exploration of co-variance for the identified features could be explored to eliminate some more features.

Also, the KNN model should be verified against a different dataset, like the dataset gnerated at the same time as the madelon dataset.