opisourml

KTR

June 2022

to jest do wklejenia do aktykułu jak dostane angielską wersje xd

1 OurML

We've decided to make an AutoMl classifier similar to flaml. The whole model can be described as a combination of following elements:

- preprocessing
- training models and selection of the best ones

1.1 preprocessing

To assert that all models we decide to train and test in the process of fitting our estimator we provided preprocessing, whitch should work for most datasets and give satisfactory results.
The following steps are:

- removal of features with zero variance
- imputation of missing data with the most common value
- substitution of values beyond .975 quantile and .025 with the threshold value
- quantile transform to make the distribution of values more reseble uniform distribution
- scaling the data with mean and variance
- for categorical variables:
  - grouping rare labels to form bigger groups
  - encoding labels with WoE

1.2 model selection

Out estimator same as flaml has a predefined selection of estimators and corresponding to them parameter search spaces. It performs random search over the hyperparameter search spaces on all models and selects the best one according to provided metric. This is a very crude and simplified copy of flamls algorithm which performs a much more complicated search and provides a way to significantly reduce training time.

1.3 performance

Due to the way a search is performed, number of models, size of hyperparameter search spaces fit time of our estimator is significant. A way to improve on that would be to further refine the search spaces, use models, which train faster or provide low computational cost default values to use instead of searching the whole space with each trainig.