Machine Learning

Regression task for Exercise 2

13.12.2021

Group 33

Ouassim Kiassa | Luis Kolb | David Siegmund

Content

- Dataset Characteristics
- Pre-processing
- Models and Approaches
- Results for our implementation
- Possible Improvements
- Comparison to other techniques
- Discussion

Dataset Characteristics

	1) Breast Cancer	2) Concrete	3) Seoul Bike Sharing
SIZE	286	1030	8760
FEATURES	30	9	14
TARGET	diagnosis (binary)	concrete strength (numerical)	no. rented bike (ordinal)

Pre-processing I

1. Scaling
   a. all continuous numerical values were scaled with “RobustScaler” (robust to outliers, quartile range: 0.25 - 0.75)
   b. Dates were scaled with an OrdinalScaler
2. Encoding
   a. multi-feature categorical data was encoded with OneHotEncoder
   b. binary (categorical) data was converted into binary numerical data

Pre-processing II

3. Dimensionality reduction
   a. the datasets were decomposed using PCA
   b. The number of attributes was incrementally decreased until the *cumulative explained variance* fell below 90 % of the explained variance achieved.

Models and Approaches

ElasticNet

DecisionTreeRegressor

> RandomForestRegressor was also tried, tended to perform even better for much longer fitting/run times

MLPRegressor

Results for our implementation

	concrete	breast cancer	seoul bike sharing
estimator	MultiLayerPerceptron	DecisionTree	MultiLayerPerceptron
parameters	{'alpha': 0.01, 'n_iter_no_change': 8, 'max_iter': 400}	default sklearn parameters	{'alpha': 0.01, 'n_iter_no_change': 8, 'max_iter': 400}
r2 score	0.71	0.74	0.60

runtime: ~30 minutes without cross-validation

Possible Improvements

- run it again using auto_ml_with_cv() in the notebook > omitted due to time limitations
- restructure stepping algorithm to accommodate string-parameters/tuples, more flexibility
- improve diagnostic tooling for debugging, break the code into even more functions to be more testable
- the PCA has to be done on the test split solely, not on the whole dataset

Lessons Learned

AutoML is hard to get right

long iteration times make development tough

saves time in the long run

existing libraries should be used if available and feasible (auto-sklearn/TPOT)

Comparison to other techniques

We compared our results to TpotAutoML.

- Run with defaults parameters for each dataset.
- run with custom parameters specifying Algorithm and hyperparameters.

Comparison to other techniques

Defaults parameters:

Pros

-gives wide range of parameters and algorithms

-explore wider combinations than homebrew pipelines

Cons

- compute intensive

-take too much time

Discussion

	concrete	breast cancer	seoul bike
r2 tpot	0.80	0.87	0.63
r2 from our own implementation	0.71	0.74	0.60