HyperparameterDB

Vaishnavi Malireddy, Nupur

Abstract: Hyperparameters are given, prior training of any model. They are essential and are useful in tweaking the model. Model behavior is depended on tuning. But selecting the values for the tuning is very long process. Our aim is to run H2O AutoML for the various runtimes and models which are generated are stored in a database. This database can be accessed by public for exploring the optimal values for any kind of algorithm.

Keyword – Hyperparameter, H2O, AutoML, Regression

1.Introduction

1.1 Background

Parameters are distinguished in two ways. Model parameters are used by the model to predict test data. While hyperparameters are specified externally.

Hyperparameters are the very important during model training as they have so much impact on model behavior. There is no procedure to select the hyperparameter values. They are picked randomly and train the model. This is repeated until the model gives good results which is a very hectic task for data scientists. Hence a database should be made that has ideal hyperparameters for all kind of algorithms. So, the purpose of this project to find the best parameters for the algorithms which are generated using AutoML for various runtimes.

1.2 Dataset

The dataset contains information on animation series. It has a target variable which is continuous, called score whose value is derived based on series popularity, rank etc.

Variables:

i) Title – title of the animation series in Japan
ii) Title_english - Title in English version
iii) Image_url – url of image
iv) Type – type of series
v) Source – weather the series is Manga or Novel or Original.
vi) Episodes – how many episodes in each series.
vii) Status – if the series is still airing or finishes airing.
viii) Airing – if its airing or not
ix) Aired_string – date when the series aired
x) Aired – duration of series
xi) Duration – duration of each episodes

xii) score – score of the series which is the target.
xiii) rank – rank of the series

2 Process

After data set cleaning is done, we have divided the data into X and Y, where X has all the column values except target and y has target column. Later, we have chosen five different runtimes – 500,700,1000,1300,1500 and ran H2O AutoML for each runtime.

H2O is a software which is used in machine learning and it is very efficient and fast. AutoML is a part of H2O which considers all the models for the given data and checks with all values of hyperparameters and gives best ones for each model. For each AutoML, different models are generated, these are saved in csv file and parameters for each model are saved in JSON file. A file called meta data is created which contains all the functionalities of data such as execution time, run id, variables of X and y etc. This is used to see the basic functionalities of the AutoML model generated for each runtime. So, leaderboard, metadata and model parameters are saved, and a database is created to generate a CSV file of hyperparameters.

3 Analysis

For each type of model, csv files are generated from database with all the models and their values of hyperparameters. These are analyzed to find three things.

3.1 Important hyperparameters

Grid search is one of the methods to find important hyperparameters. Values of hyperparameters are kept constant except one and are passed to Grid Search for finding RMSE. This particular parameter is considered important if there are variations in the RMSE values. This process is repeated until all parameters are passed by keeping others constant.

3.2 Range of hyperparameters

Range is calculated by finding minimum and maximum values for hyperparameters in all algorithms.

3.3 Compare the range of values across models for different hyperparameters

This is done by finding hyperparameters which are common in all the models and the range of the values attained can be compared with all the models attained for different run times.

4 Results

For each runtime, 3 to 5 types of algorithms were created.

RUN TIME (sec)	Algorithms generated	Total number of models
333	StackedEnsemble, Xgboost,GLM	6
500	Staked Ensemble, Xgboost, DRF,GBM, GLM, XRT	9
700	Staked Ensemble, Xgboost, DRF,GBM, GLM, XRT	11
1000	Staked Ensemble, Xgboost, DRF,GBM, GLM, XRT	18
1300	Staked Ensemble, Xgboost, DRF,GBM, GLM, XRT	29
1500	Staked Ensemble, Xgboost, DRF,GBM, GLM, XRT	33

Staked Ensemble comprises of all the models. Hence, it is not considered for finding the hyperparameters and instead all the models are taken individually and correspondingly hyperparameters are found.

Analysis:

Following are the important hyperparameters found:
For XGBoost - max_depth, ntrees, min_rows
For GBM - sample_rate for GBM.
For GLM, XRT, DRF – nothing

The ranges for each algorithm are as follows

XGBoost:

Hyperparameters	Minimum	Maximum
nfolds	5	5
seed	-8.186597997308668e+18	8.50284134061186e+18
ntrees	13.0	276.0
max_runtime_secs	0.0	53.0
stopping_rounds	0.0	3.0
max_depth	5.0	20.0
learn_rate	0.05	0.05
min_rows	0.01	20.0
sample_rate	0.6	1.0
col_sam_rate	0.6	1.0
col_sam_r_per_tree	0.7	1.0
score_tree_interval	5.0	5.0

GBM:

Hyperparameters	Minimum	Maximum
nfolds	5	5
max_depth	3.0	16.0
min_rows	3.0	16.0
stopping_tolerance	0.008312003	0.008312003
seed	-9e+18	8.93e+18
sample_rate	0.5	1.0

GLM:

Hyperparameters	Minimum	Maximum
nfolds	5	5
seed	-5.91e+18	6.33e+18
max_iterations	300.0	300.0
objective_epsilon	300.0	300.0
gradient_epsilon	1e-06	1e-06
lambda_min_ratio	0.0001	0.0001
obj_reg	6.91e-05	6.91e-05

XRT:

Hyperparameters	Minimum	Maximum
nfolds	5	5
seed	-7.04e+18	6.82e+18
ntrees	14.0	50.0
stopping_tolerance	0.008312003	0.008312003

DRF:

Hyperparameters	Minimum	Maximum
nfolds	5	5
stopping_tolerance	0.008312003	0.008312003
seed	-6.25e+18	8.56e+18

For the above models, the common hyperparameters are found to be nfolds and seed. nfolds value is 5 for all algorithm and seed value differed from -9e+18 to 8.93e+18.

5 Conclusion

Hyperparameters play an important role for tuning the models. Hence, a database is created containing many models with n number of runtimes. Further, many more hyperparameters will be generated and stored in database. And to make the database easily accessible, a website will be created such that any user can easily get the values of hyperparameters for given runtime.

References

https://github.com/prabhuSub/Hyperparamter-Samples

https://github.com/nikbearbrown/CSYE_7245/tree/master/H2O

https://dzone.com/articles/exploring-h2oai-automl

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html

http://docs.h2o.ai/h2o/latest-stable/h2o-docs/grid-search.html