HyperparameterDB

Urja Jain[1], Prakruthi Bagur Suryanarayanaprasad[1]

[1]Northeastern University, Boston MA
{jain.u , bagursuryanarayana.p}@husky.neu.edu

ABSTRACT
In the field of data science Hyperparameters are very much important also their values are very crucial. They are responsible for controlling training behaviour in-order to increase accuracy of prediction outcomes. Determining hyperparameters is a tedious and cumbersome task. The main goal of this research focuses on determining the important hyperparameters with their values for proper tuning. To achieve this, this research leveraged H2O.ai's python module python called H2O, gives flexibility to run the process for numerous run times. Models are generated by running H2OAutoML for various runtimes (300s, 500s, 700s, 1000s, and 1200s).

Key Words – Hyperparameters, H20AutoML, H2O, predict, accurate.

1. INTRODUCTION

1.1 Background

The parameters which are defined before training any machine learning algorithm are called **Hyperparameters**. In other sense, these are fixed parameters, whose values are predefined before training any algorithm. Significant role is played by hyperparameters when tuning the algorithms thus they are called as tuning parameters. Selecting the good hyperparameter for the purpose of tuning is one of the major aspects in Data Science. It provides two benefits: firstly, efficiently searching space of all possible hyperparameters. Secondly, it's easy to manage the experiments to tune the large dataset for tuning the hyperparameter. Challenge with hyperparameters is, they don't have that particular value which will work everywhere as desired. Further, hyperparameters can be separated into two categories: optimizer hyperparameters and Model specific hyperparameters. Former one relates to process of training and optimization. Few examples of optimizer hyperparameters are learning rate, minibatch size number of epochs. Amongst these learning rate seems to be the most important one which must be tuned. Model hyperparameters are related with the structure of the model. This research solely based on determining the hyperparameters for users to visualize and understand hyperparameters, which should be considered in-order to predictive power of their models. This research is a task based research, from selecting data performing data cleaning to finding the significant hyperparameters and their ranges using H2O python module. In this research we got GLM, GBM, DRF, XRT algorithms generated by running H2O.

1.2 Data Set Overview

The dataset used in this research is taken from the kaggle data. This data is related to the employee access. The basic requirements of the data set to predict the whether the employee is given access or not. That means it has two classes as approval or denial which makes this data to be Classification type. The dependent

variable "ACTION" is of binary nature. There are around 9 variables that are independent in the dataset. The dataset is ran on 5 different runtime using H2O.

Variables:

i. ACTION: Access approve (1) or Denied(0)
ii. RESOURCE: unique resource ID
iii. MGR_ID: Manager's employee ID
iv. ROLE_ROLLUP_1: Role grouping category ID = 1 for the company
v. ROLE_ROLEUP_2: Role grouping category ID = 2 for the company
vi. ROLE_DEPTNAME: role department description
vii. ROLE_TITLE: Title description of business role
viii. ROLE_FAMILY_DESC: Extended description of role family
ix. ROLE_FAMILY: Description of role family
x. ROLE_CODE: Unique role code

1.3 Classification Type problem

These are the class based; basically they are categorical type problems. If the dependent variable is in the form of binary or multiclass then they are said to be classification type problem. Generally evaluation metric(s) used for these types of problems like ROC, AUC, logloss, multiclass loss etc. If AUC is closer to 1 and logloss and other error term is closer to 0 then it will yield a good model.

1.4 Data cleaning

Data cleaning is very much important for predicting better results. Proper data cleaning is required before proceeding to any analysis, if data is not cleaned then the outcomes might deviate to a large extent from the desired result. For this research we did the data cleaning by checking the null values and the data types of the variables. In this research, the data set containing the hyperparameter values obtained for different algorithms by running H2O required data cleaning to do further analysis. The data types of the variables were checked and converted into the data type suitable for the analysis (object data type to numeric).

1.5 Process

After proper data cleaning multiple processes were sequentially implemented in order to find the important hyperparameters for best models, their ranges and compared their ranges across different algorithms. The processes implemented are:

i. Memory allocation and meta data creation for storing the details required for the execution of the analysis.
ii. Initialized H2O and used AutoML for determining the hyperparameters values generated by models at different run times
iii. Saved the metadata, models and their hyperparameters to different JSON/ CSV files.

1.6 Analysis of Hyperparameters

Using the data in the above files, further analysis was done to find important hyperparameter, their ranges and comparison of the ranges across different algorithms.

- H2O grid search was used to find important hyperparameters. For each algorithm a specific classifier is used to run the grid search. Random hyperparameters from each algorithm was chosen, to observe its impact on score generated. This process is repeated with each hyperparameter and compared the results to find the important hyperparameter.
- Next part of analysis was to determine the range of hyperparameter values.
- These ranges are compared across different algorithms.

Result

Summary Table

RUN TIME (sec)	Algorithms generated	Total number of models
300	GBM, GLM, DRF, XRT, DeepLearning	11
500	XRT, DRF, GLM	6
700	GBM, GLM, DRF, DeepLearning	22
1000	GBM, GLM, XRT, DeepLearning	25
1200	GBM, GLM, DRF, XRT, DeepLearning	29

It can be referred from the summary table that as the runtime increases the number of models generation will increase.

Analysis part

Important hyperparameters

Used H2O grid Search

Model	Hyperparameter
XRT	mtries
DRF	mtries
GLM	alpha
GBM	learning_rate

Range of Hyperparameters:

GBM

Hyperparameter	Minimum	Maximum
learn_rate	0.001	0.8
col_sample_rate	0.4	1.0
tweedie	1.5	1.5
quantile_alpha	0.5	0.5
huber_alpha	0.9	0.9
sample_rate	0.5	1.0
col_sample_rate_per_tree	0.4	1.0
min_split_improvement	1e-05	0.0001
max_abs_leafnode_pred	1.7976931348623157e+308	1.7976931348623157e+308

DRF

Hyperparameter	Minimum	Maximum
mtries	-1	-1

XRT

Hyperparameter	Minimum	Maximum
mtries	-1	-1

GLM

Hyperparameter	Minimum	Maximum
seed	711766579752531550	7150684335027209933
tweedie_variance_power	0.0	0.0
tweedie_link_power	1.0	1.0
alpha	0.0	1.0
lambda	0.0013175999248032143	0.40044436071981493
theta	1e-10	1e-10

Comparing the ranges of Hyperparameters:

Model	Hyperparameter	range
DRF	mtries	-1
XRT	mtries	-1

Conclusion

Based on the results achieved, it is concluded that there are certain hyperparameters that play a major role in tuning models. The hyperparameter values chosen help us to make the models function more efficiently by getting improved score of metrics. For future scope, this research will help users by giving the values hyperparameters for the purpose of tuning, beforehand to improve the efficiency of the models.

Results depict that, alpha, learning_rate and mtries are the important hyperparameters to tune GLM, GBM and XRT/DRF algorithms. XRT and DRF algorithms have common hyperparameters and ranges. The range of each hyperparameter can be utilized to further tune the models for different and unknown datasets with classification type problem.

References

[1]https://github.com/prabhuSub/Hyperparamter-Samples

[2]https://github.com/nikbearbrown/CSYE_7245/tree/master/H2O

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

[4]https://github.com/AmiGandhi/Top-1-percent-in-Kaggle-Competition

[5]https://towardsdatascience.com/grid-search-for-model-tuning-3319b259367e