HYPERPARAMETER SEARCH

Hyperparameter-db-project-db15

Pallavi Kashyap | INFO6210 | April 26,2019

ABSTRACT

* Hyperparameter tuning is choosing a set of optimal hyperparameters for a learning algorithm. Hyperparameter tuning is critical for building accurate models, yet most researchers use personal experience to specify their range.

There are two different methods for optimizing hyperparameters:

Grid Search

Random Search

## GOAL

This project is to build the database from thousands of public classification and regression dataset using the Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM) and XGBoost algorithms. This data will allow us to estimate the relative importance of hyperparameters, appropriate hyperparameter ranges and to build predictive models of hyperparameter values. Out of those thousands, I am only working on one Dataset-Adult.csv from kaggle.

Description of the Dataset:

This data was extracted from the census bureau database found at

--http://www.census.gov/ftp/pub/DES/www/welcome.html

-- Donor: Ronny Kohavi and Barry Becker,

-- Data Mining and Visualization

-- Silicon Graphics.

--Split into train-test (06, 0.85, 0.7random).

--48842 instances, mix of continuous and discrete (train=32561, test=16281)

--45222 if instances with unknown values are removed (train=30162, test=15060)

--Duplicate or conflicting instances: 6

-- Probability for the label '>50K' : 23.93% / 24.78% (without unknowns)

--Probability for the label '<=50K' : 76.07% / 75.22% (without unknowns)

References:

https://towardsdatascience.com/hyperparameter-tuning-c5619e7e6624

https://www.kaggle.com/wenruliu/adult-income-dataset/version/2?#adult.csv

https://github.com/prabhuSub/Hyperparamter-Samples/blob/master/Meta\_Data\_Generated/StudentPerformanceDataset/StudentPerformance\_100.ipynb

https://github.com/nikbearbrown/CSYE\_7245/blob/master/H2O/H2O\_automl\_model.ipynb

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