INFO 6210 Project Proposal
Modeling Hyperparameter Database

Dhawal Priyadarshi, Mansi Nagraj, Mayur Vyas

Background

A model hyperparameter is a configuration that is external to the model and is often used in processes to help estimate model parameters. They are usually tuned by the practitioner for a given predictive modeling problem. [3]
In machine learning scenarios, a significant part of model performance depends on the hyperparameter values selected. The goal of hyperparameter exploration is to search across various hyperparameter configurations to find the one that results in the optimal performance. [2]
The hyperparameter database to be developed as a part of this project is an open resource with algorithms, tools, and data that allows users to visualize and understand how to choose hyperparameters that maximize the predictive power of their models. Phase I of the project involves gathering the data in json files containing predicted target variables, hyperparameters, meta-data etc. by running different models (with varying hyperparameters) on a single dataset using H2O.
Currently, the hyperparameter database analyzes the effect of hyperparameters on the following algorithms: Distributed Random Forest (DRF), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM). Naïve Bayes Classifier, Stacked Ensembles, Xgboost and Deep Learning Models (Neural Networks). [1]

Objectives

Typically, hyperparameter-tuning is painstakingly manual, given that the search space is vast and evaluation of each hyperparameter configuration expensive. The correct choice of hyperparameters plays a crucial role in creating the best ML models besides training those on proper datasets. Moreover, certain hyperparameter configurations tend to be common for common ML use-cases.

Therefore, this project aims at creating the necessary database infrastructure that makes the storage and retrieval of relevant datasets and hyperparameters easy for the user (the model creator). This includes curating *datasets* for modeling and organizing *hyperparameter metadata* in a fully normalized (3NF) and optimized database with related functions and procedures.

Dataset

The project deals with obtaining (downloading/scraping) training datasets for the machine learning models in a curated manner such that these are easy to retrieve for specific model training purposes. Curation details include information pertaining to the datasets like the size of the data, the type of data (classification/regression), the number of times these have been used in training, the kind of models that have been created using these, etc.

Hyperparameter metadata set

These are the output storage sets and related tables around hyperparameter metadata sets derived after running the model. This part deals with ingesting the json files generated from the model runs and setting them up in fixed schemas in the database, for easy retrieval and insight generation.

Functions & stored Procedures

These help the user in obtaining useful information from the database quickly. Furthermore, some functions and stored procedures could be used to obtain performance-based information such as the number of times a dataset has been used for training, various models trained using these datasets, average errors and accuracy associated with each model, best-fit model and the corresponding best hyperparameter value.

Proposed Actions

Data Gathering

This involves gathering 2 to 4 datasets of different "types" of data used in training ML models. Different types could include datasets for regression modeling, classification modeling, etc. The dataset will be fed to H2O to generate JSON files containing hyperparameter data.

Conceptual Data Model

Conceptual Data Model is built to represent the architecture of the database. This includes deciding the entities, the relational entities, and associated attributes.

ER Diagram & Physical Model

Next Entity-Resolution diagrams are created corresponding to the conceptual model. This involves multiple iterations on the ER to add/modify data entities and create a fully normalized (3NF) structure. After iterating over ER diagrams, the database will be realized physically in the database as database schemas.

Parsing and Data-entry

Creating scripts that parse the JSON data into neat CSV files that can be imported into the database. Additionally, a data pipeline is created that automatically updates the database as new data is being produced.

Optimization

Indexing and minor structural changes to improve database performance.

Report

Preparing a final summary report including the details of all the artifacts to be used as a reference for database users.

Future Scope

It is imperative that the database be scaled to millions of datasets and related hyperparameter information. Also, the database backend could be enhanced with a front-end that could help leverage its power through dynamic dashboard providing useful insights into datasets, models and hyperparameters that are relevant to the end-user.

Moreover, the hyperparameter database should be used to get insights and recommendations from. For example, the DB could be used to build ML models that could predict hyperparameters, visualizing and teaching statistical concepts such as power and bias/variance tradeoff, etc.

References

[1]: https://github.com/skunkworksneu/Projects/blob/master/Hyperparameter%20Database.pdf
[2]: https://www.analyticsindiamag.com/
[3]: https://machinelearningmastery.com/