AI Skunkworks Project

Hyperparameter Database

Project Manager: Prabhu Subramanian (subramanian.pr@husky.neu.edu)

Co-Project Manager: Chitra Paryani (paryani.c@husky.neu.edu),

Ami Gandhi (gandhi.am@husky.neu.edu)

Expected Start: March 16th, 2019

Expected Project Duration: 3 months (TBD (For Spring 2019 students who need to submit a project, the deadline for tasks will be course project deadline)

Abstract:

Hyperparameters are parameters that are specified prior to running machine learning algorithms that have a large effect on the predictive power of statistical models. Knowledge of the relative importance of a hyperparameter to an algorithm and its range of values is crucial to hyperparameter tuning and creating effective models.

The hyperparameter database is a public 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.

The hyperparameter database is created by running millions of hyperparameter values, over thousands of public datasets and calculating the individual conditional expectation of every hyperparameter on the quality of a model.

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).

The hyperparameter database also uses these data to build models that can predict hyperparameters without search and for visualizing and teaching statistical concepts such as power and bias/variance tradeoff.

Project Details:

The project will be divided in 2 teams –

1. Generating models using H2O software to find best hyperparameters (Data Science)
2. Creating a conceptual model and storing all the data into a physical database (DMDD)

Maximum 4 people allowed in both teams (max 2 for data science part and max 2 for database). And this project can be taken up by any number of people for different datasets.

Prerequisites:

Able to conduct basic machine learning tasks (Data Science)

Basic knowledge of SQL and database management system (Database)

Deliverables:

In hyperparameter project, the deliverables are to gather a list of data sets, type of datasets, and hyperparameters by running an expanded list of datasets. All this information will be embedded in a database management system, to be incorporated into a website where it is easy to be searched and used by the public.

Tasks for Data Science team:

1. Pick up a unique data set from any source e.g. Kaggle, UCI machine learning, etc.
2. Identify the type of dataset – Regression/Classification
3. Perform data cleaning and pre-processing
4. Run H2O to get the best models via leaderboard
5. Get the best hyperparameters for the best model on leaderboard
6. Save the model in a file to be embedded in the database management system
7. Repeat all above steps for different run times

Tasks for Database team:

1. Create a conceptual model of the database
2. Create an E-R diagram
3. Perform database normalization – 1NF, 2NF, 3NF
4. Create tables as per the requirement
5. Insert data into the respective tables from the file which is generated and stored by the Data Science member (step 6 above)
6. Write standard queries/use-cases (minimum 10)
7. Perform analytics on the database using queries e.g. getting the best value for the hyperparameter from the database

Project Deadline:

The project needs to be coordinated in a team of 2 and finish it by the given deadline.