**Anushadevi Mohan**

**May 11, 2016**

## Topic: Machine Learning with H2O.ai

## Concentration: Protein tertiary structure prediction

**Problem/Purpose:** The goal of this project is twofold.

1. Evaluate the following different ways of using H2O.ai for Machine learning
   1. H2O directly from Python
   2. H2O directly from R
   3. PySparkling from local standalone Spark cluster
   4. H2O Flow on Desktop mode
2. Demonstrate the use of following H2O algorithms on a lesser known Benchmark dataset used by academics and compare the performance metrics
3. Deep Learning
4. Gradient Boosting Machine
5. Generalized Linear Model
6. Distributed Random Forest
7. Ensembles stacking using R

**Big Data Set:**

data files: foldpro.dataset

description of data: The Protein Fold recognition is a posed as binary classification with a goal of identify proteins that have similar tertiary structure. The FOLDpro dataset is a comprehensive dataset used to benchmark multiple algorithms for this task. The total number of records in the dataset is 951,600 ( i.e 976 x 975 ). Each record consists of 84 features.The file is in LIBSVM format

source data URL: <http://mine5.ics.uci.edu:1026/fold_help.html>

Size(uncompressed): 1.5GB

Format of data file: csv

**Hardware:**

* Ubuntu 15.04 on a workstation (Dual CPU + 48 GB memory)

**Software:**

|  |  |
| --- | --- |
| **Technology/tools** | **Description** |
| H2O Flow | Web based interactive environment to perform Machine Learning |
| H2O on Python | One of H2O platforms for machine learning |
| H2O on R | One of H2O platforms for machine learning |
| H2O on Spark | One of H2O platforms for machine learning |

**Overview of steps:**

Install h2o in python, R, and Spark & verifying installation for all the platforms

Install h2oensemble R package

Install and verify installation of H2O flow

Download and Install executables for benchmarking performance of algorithms as related to the problem statement of this project.

Run aforementioned machine learning algorithms on foldpro dataset and compare performance metrics for the same

**Summary:** The classifiers Gradient Boosting, Distributed Random Forest, Generalized Linear modeling, Deep learning and Ensemble, which is a combination of classifiers were trained on this dataset. And as expected the Ensemble of classifiers performed better than the individual classifiers to predict whether or not two proteins share a similar tertiary structure

**YouTube URLs here:**

2 min video: <https://www.youtube.com/watch?v=v2yGoV2HDTc>

15min video: <https://www.youtube.com/watch?v=-q4orm8zFDQ>