# Predicting the Running Time of Sklearn Programs.

## Efficiently running distributed applications on shared infrastructure is challenging. Past research shows that choosing proper hardware configuration can significantly improve performance and reduce cost. To choose optimal configuration, predicting the performance of different applications accurately is required.

We consider a simplified version of this challenge: Predicting the performance of machine learning programs on a single server. Scikit-learn (sklearn) is a machine learning library for python. The dataset provided describes a few examples of running SGDClassifier in sklearn. The features of the dataset describe the SGDClassifier as well as the features used to generate the synthetic training data. The data to be analyzed is the training time of the SGDClassifier.

Together with this document you can find a sample training dataset (“sample\_train.csv”).

'Time' is the training time of the model. Specifically,'n\_samples','n\_features' describes how the synthetic dataset for training is generated using [sklearn.datasets.make\_classification](http://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.html), after generating dataset, we make classification with SGDClassifier, and'l1\_ratio','alpha','max\_iter' describes the setup of [sklearn.linear\_model.SGDClassifier](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html" \l "sklearn.linear_model.SGDClassifier). All the feature names follows the definition in sklearn documentation.

The goal of this project is to minimize the error of predicted running time, i.e., you should predict the running time of a testing set that is close to the ground-truth running time. The evaluation metric used is mean-squared error. You can find “sample\_test.csv” and “sample\_submission.csv” for the testing set and solution submission format.

The project consists of 2 phases:

* Phase 1: Prior to your visit at HKUST, you can explore the sample dataset and draft solutions. You are suggested to work on the dataset and prepare models in advance.
* Phase 2: During your visit, we will provide you a larger dataset with more samples and features. You can then work on your solution with the expanded dataset. Please notice that the data we provide in phase 1 is only an example, an accurate model on the example dataset doesn’t necessarily mean good performance on real dataset at this phase.

Other information and requirements for the project:

* All the training and testing datasets are generated on the same server.
* No restrictions on the programming language or machine learning techniques used. However, *you are not allowed to run the sklearn programs to expand training set or predict testing set*.
* You need to program and train your model on your own laptop computer.

If you have further enquiries, feel free to drop an email to Jiacheng Xia at jxiaab {at} ust {dot} hk.