Faculty of Arts & Sciences
Department of Computer Science
CMPS 261 – Machine Learning
Spring 2023 – Course Project
Due April 18th, 2023 11:59 PM

Problem Statement

The field of high-energy physics is devoted to the study of the elementary constituents of matter. By investigating the structure of matter and the laws that govern its interactions, this field strives to discover the fundamental properties of the physical universe. The primary tools of experimental high-energy physicists are modern accelerators, which collide protons and/or antiprotons to create exotic particles that occur only at extremely high-energy densities. Observing these particles and measuring their properties may yield critical insights about the very nature of matter. Collisions at high-energy particle colliders are a traditionally fruitful source of exotic particle discoveries. Finding these rare particles requires solving difficult signal-versus-background classification problems, hence machine-learning approaches are often used. Given the limited quantity and expensive nature of the data, improvements in analytical tools directly boost particle discovery potential.

The vast majority of particle collisions do not produce exotic particles. For example, though the Large Hadron Collider (LHC) produces approximately 10^{11} collisions per hour, approximately 300 of these collisions result in a Higgs boson, on average. Therefore, good data analysis depends on distinguishing collisions which produce particles of interest (signal) from those producing other particles (background).

In this project, your task is a classification problem where the goal is to distinguish between a signal process where new theoretical Higgs bosons (HIGGS) are produced, and a background process with the identical decay products but distinct kinematic features.

The above description of the problem is formed of snippets collected from the original paper introducing the problem:
https://www.nature.com/articles/ncomms5308

Dataset

Each process (signal or background) in the dataset is represented by 28 features. The first 21 features (columns 2-22) are kinematic properties measured by the particle detectors in the accelerator. The last seven features are functions of the first 21 features; these are high-level features derived by physicists to help discriminate between the two classes. There is an interest in using deep learning methods to obviate the need for physicists to manually develop such features.

The first column is the class label (1 for signal, 0 for background), followed by the 28 features (21 low-level features then 7 high-level features): lepton pT, lepton eta, lepton phi, missing energy magnitude, missing energy phi, jet 1 pt, jet 1 eta, jet 1 phi, jet 1 b-tag, jet 2 pt, jet 2 eta, jet 2 phi, jet 2 b-tag, jet 3 pt, jet 3 eta, jet 3 phi, jet 3 b-tag, jet 4 pt, jet 4 eta, jet 4 phi, jet 4 b-tag, m_jj, m_jjj, m_lv, m_jlv, m_bb, m_wbb, m_wwbb. For more detailed information about each feature see the original paper.

You are provided with the file HIGGS_train.csv which consists of 600,000 training examples. You will not gain access to the testing dataset that will be used to evaluate your model.

Scoring

Your model will be evaluated on a hidden dataset, using the Accuracy metric.

Faculty of Arts & Sciences
Department of Computer Science
CMPS 261 – Machine Learning
Spring 2023 – Course Project
Due April 18th, 2023 11:59 PM

Deliverables

In addition to submitting all the required files through Moodle, you are required to maintain a Github repository with a properly documented code and include the link of that repository in your submission. You are also expected to submit a report detailing your work, including any pre-processing steps performed on the data, all experiments you performed during the training phase, and how you validated your methods. You will be expected to present the report orally. You are also expected to submit a saved machine learning model that can be used for inference to be evaluated for generalization on new data.

Grading

Your grade will include the Github repository, code, documentation, report, and presentation. In the spirit of machine learning competitions, bonus points will be given as awards to the 3 top scoring teams.

In addition to your GitHub submission, add all relevant files (e.g. code, report) to one folder: groupname.higgs_boson.zip (or .rar) and submit it to moodle.