Classification Algorithms on GPU Variants

TEAM 2: Kevin Portillo, Taylar Stowers, Kimberley Davis

September 24, 2018

Motivation

Machine Learning (ML) has become a hot topic in the recent years. Various industries have begun to dive into ML concepts and algorithms to improve decision making, predict market sentiment, and understand consumers. College undergraduates and interested programmers have noticed this trend and have begun diving into ML techniques and concepts. Through decades of testing, it has been shown the Central Processing Units (CPUs) alone are not suffice to run ML algorithms efficiently, even if the Random Access Memory (RAM) cards are upgraded. Graphical Processing Units (GPUs) were thus found to be a parameter whithin a system that could yield better results. In recent years, GPUs have be proven to be useful tools when attempting to implement ML in a project. A large number of GPUs within a system, known as a GPU cluster, have been contributing to the efficiency of super computers for decades.

Problem

For this project, we propose to run two ML classification algorithms (CAs) on a single dataset: Support Vector Machine (SVM) and K-Nearest Neighbor (kNN). The goal is to determine which system can perform ML classification the best at a reasonable price so that novice programmers can use as reference. Intuitivelty, the platform with the most amount of GPUs will preform the best but at its price, will not suit well for novice students on a budget. With this in mind we propose running the CAs on several platforms: a laptop with a dedicated GPU, a desktop with one GPU, another desktop with three GPUs, another with six GPUs, finally with the help of the Center for Advanced Computing and Data Science (CACDS) we will run these CAs on the Sabine Cluster that host a total of 5704 CPU cores and 12 GPU nodes.

Resolution

The two CAs that will be used to test each system will be SVM and kNN. The dataset has yet to be determined but we are looking for one with a few million tuples. The CAs will be implemented in Python using Scikit-learn, a free software ML library for the Python programming language (PL). The platforms on which we will test our algorithms will be provided by one of our team members who owns the laptop with dedidicated GPU and the single, triple, and sextuple GPU desktops. The CACDS intitute will provide access to the Sabine Cluster as mentioned above. Bash scripting will also be used to consolidate the testing procedures and be the intermediary between the user and the Operating System (OS).