Ethics Pledge

Consistent with the above statements, all homework exercises, tests and exams that are designated as individual assignments MUST contain the following signed statement before they can be accepted for grading.

I pledge on my honor that I have not given or received any unauthorized assistance on this assignment/examination. I further pledge that I have not copied any material from a book, article, the Internet or any other source except where I have expressly cited the source.

Signature: Haodong Zhao Date: Apr 10th. 2019

Please note that assignments in this class may be submitted to

www.turnitin.com, a web-based anti-plagiarism system, for an evaluation of their originality.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Reading review**

**XGBoost: A scalable Tree Boosting System**

Tree Boosting is an efficient and widely used machine learning method. XGBoost is a scalable end-to-end tree boosting system that is widely used by data scientists to obtain state-of-the-art results in many machine learning challenges.

The most important factor in the success of XGBoost is its extensibility in all scenarios. The system runs more than ten times faster on a single machine than existing popular solutions.

**Split finding algorithms:**

1. Basic exact greedy algorithm
2. Approximate algorithm
3. Weighted quantile sketch
4. Sparsity-aware split finding

**system design:**

1. Column block for parallel learning: To reduce the cost of sorting, store the data in a block. The column block structure also supports column subsampling, as it is easy to select a subset of columns in a block
2. Cache-aware Access: cache-aware implementation of the exact greedy algorithm runs twice as fast as the naive version when the dataset is large
3. Blocks for Out-of-core Computation

This article proposed a novel sparsity aware algorithm for handling sparse data and a theoretically justified weighted quantile sketch for approximate learning.

This article also describes the cache access mode, where data compression and slicing are essential elements for building an extensible end-to-end system for tree promotion. XGBoost is able to solve real-world problems with a minimum amount of resources.