CSE 234 Fall 2021 Midterm Exam

Full Name:

Student ID:

INSTRUCTIONS
1. This exam is for 75pts (+ 5pts extra credit). You have 80min.
2. You can have up to one sheet of notes with you. Apart from that, this exam is closed books/notes/electronics/peers.
3. Please wait until being told to start reading and working on the exam.
4. Make sure that your writing is clear and legible!
5. Ask for extra white sheets for rough work if you need any.

<table>
<thead>
<tr>
<th></th>
<th>Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q 1</td>
<td>48</td>
<td></td>
</tr>
<tr>
<td>Q 2</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Q 3</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Q 4</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Q 5</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>ECQ</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>75 (+5)</td>
<td></td>
</tr>
</tbody>
</table>
THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
Q1. [16 x 3pts = 48pts] For each of the following questions, select the correct option; only one is correct.

1. Which of the following model selection approaches/systems can *not* ensure logical equivalence to sequential SGD for the models trained?
   (A) Horovod
   (B) Asynchronous Parameter Server
   (C) Model hopper parallelism
   (D) Task parallelism

2. Which of the following consistency models in GraphLab typically offer(s) the maximum amount of parallelism and throughput for large-scale graph analytics?
   (A) Full consistency
   (B) Edge consistency
   (C) Vertex consistency
   (D) Eventual consistency

3. In a 2-D tradeoff plot to compare model selection approaches with final best accuracy on the x-axis and completion time on the y-axis, toward which corner is the Pareto frontier expected to lie? Note that “left” and “right” here refer to your (the viewer’s) left and right.
   (A) Top left
   (B) Top right
   (C) Bottom left
   (D) Bottom right

4. What prior ML systems approach does AWS SageMaker adopt to get the independent workers that train on streaming data to reconcile the model with each other?
   (A) Parameter Server
   (B) XGBoost
   (C) Horovod
   (D) Model Averaging

5. Which of the following ML activities is *not* part of the “model selection triple”?
   (A) Feature engineering
   (B) Algorithm selection
   (C) Data cleaning
   (D) Hyper-parameter tuning

6. Which of the following does *not* have a sequential data access pattern?
   (A) Decision tree
   (B) K-means clustering
   (C) Logistic regression with BGD
   (D) Naive Bayes
7. What is the term used in the literature for pushing ML computations through joins to avoid denormalization when learning over multi-table data?
   (A) Normalized learning  (B) Factorized learning
   (C) Push-down learning  (D) Refactored learning

8. Which function in the parallel UDA abstraction tends to become a major scalability bottleneck at very large cluster scales (1000s of workers)?
   (A) Initialize  (B) Transition
   (C) Merge  (D) Finalize

9. Which of the following is a popular model exchange format for deep nets?
   (A) ODLX  (B) ONNX
   (C) OMMX  (D) OMDX

10. Which component of test error is not affected by ensembling of decision trees?
    (A) Bias  (B) Variance
    (C) Bayes Noise  (D) None of the other options

11. Which DL model family will benefit the most from more support for “model parallelism” in DL systems?
    (A) CNNs  (B) RNNs
    (C) MLPs  (D) Transformers

12. In what way does CPU cache-aware staging of statistics during GBDT computations help reduce runtimes in XGBoost?
    (A) Reduces DRAM stalls  (B) Reduces time complexity
    (C) Reduces random reads to disk  (D) Raises instruction-level parallelism

13. Which of the following ML systems is the most integrated with the PyData stack?
(A) MADlib  
(B) Mahout  
(C) Spark ML  
(D) GraphLab  

14. Which of these layers in a memory hierarchy has the highest capacity to cost ratio?  
(A) CPU caches  
(B) GPU memory  
(C) DRAM  
(D) Flash SSDs  

15. What is the name of the new programming paradigm ushered in by DL systems?  
(A) Imperative programming  
(B) Differentiable programming  
(C) Integrative programming  
(D) Functional programming  

16. Which of the following optimizations in TVM does not focus specifically on mitigating the performance impact of memory stalls?  
(A) Operator fusion  
(B) Nested parallelism  
(C) Tensorization of operators  
(D) Pipelining of operators
Q2. [2 x 3pts] Early Stopping in Hyperband.

Suppose you run Hyperband for hyperparameter tuning of an ML classifier with the following knobs: \( R = 81; \eta = 3 \). It yields the following brackets in the same table format from the paper as explained in class:

| \( i \) | \( n_i \) | \( r_i \) | \( s = 4 \) | \( n_i \) | \( r_i \) | \( s = 3 \) | \( n_i \) | \( r_i \) | \( s = 2 \) | \( n_i \) | \( r_i \) | \( s = 1 \) | \( n_i \) | \( r_i \) | \( s = 0 \) | \( n_i \) | \( r_i \) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0  | 81 | 1  | 27 | 3  | 9  | 9  | 6  | 27 | 5  | 81 |
| 1  | 27 | 3  | 9  | 9  | 3  | 27 | 2  | 81 |   |   |
| 2  | 9  | 9  | 3  | 27 | 1  | 81 |   |   |   |   |
| 3  | 3  | 27 | 1  | 81 |   |   |   |   |   |   |
| 4  | 1  | 81 |   |   |   |   |   |   |   |   |

1. How many configs ran for at least 10 epochs?

2. How many configs got killed in all?

Q3. [2 x 3pts] Runtimes of SGD.

Suppose you use SGD model averaging to train a GLM on 4-worker cluster. Shuffling the datasets takes 10min. An epoch of distributed SGD takes 5min. To accelerate learning you decide to shuffle the dataset only once every 2 epochs.

1. What is the total runtime of training the model for 8 epochs?
2. You double the number of workers and also decide to train for more epochs. The runtime of shuffling is now only 6min; an SGD epoch, only 3min. How many additional full epochs of training can you afford to run now in the same time as in the last question?


Suppose you run a DL model selection workload on a 10-worker cluster with 20 training configurations, each trained for 25 epochs. All model sizes are roughly 100 MB each.

The dataset size is 500 GB. It is uniformly randomly sharded on the cluster. The number of data examples is 100 million. The mini-batch size for SGD is fixed to 100.

1. What is the communication cost (in bytes) of Horovod? Round to the nearest multiple of 10 PB.

2. What is the communication cost (in bytes) of MOP? Round to the nearest multiple of 10 GB.
Q5. [9pts] You are given a large tabular dataset \((Y, X)\) on HDFS with feature vector \(X\) having only a few dozen features, all numeric. You decide to statistically normalize the data before training an SVM. This involves subtracting the mean and dividing by the standard deviation for each feature, with the statistics computed over the whole data.

Write succinct pseudocode for (or explain precisely) a single MapReduce job to compute all the means and standard deviations. It should be scalable along the number of examples/rows. Note that you only need to obtain the statistics for this question, not update the features. Write only your final answer in the sections on this page and next.

(Hint: Standard deviation of \(X_i\) can be expressed as \(\sigma(X_i) = (E[X_i^2] - E[X_i]^2)^{\frac{1}{2}}\)

Sharding Scheme:

Map Function:
Reduce Function:
Extra Credit Question. [5pts] You are given a matrix $A$ represented as a relation with one tuple per cell in the following schema: $A(row, column, value)$. Write an SQL query (SELECT ...) to compute the Gramian matrix $A^T A$. Write each clause in a new line.