UC Berkeley School of Information | MIDS Program

w261 Final Project

Fall 2018

Overview

Throughout this course you've engaged with key principles required to develop scalable machine learning analyses for structured and unstructured data. Working in Hadoop Streaming and Spark you've learned to translate common machine learning algorithms into Map-Reduce style implementations. You've developed the ability to evaluate Machine Learning approaches both in terms of their predictive performance as well as their scalability. For the final project you will demonstrate these skills by solving a machine learning challenge on a new dataset. Your job is to perform Click Through Rate prediction on a large dataset of Criteo advertising data made public as part of a Kaggle competition a few years back. As you perform your analysis, keep in mind that we are not grading you on the final performance of your model or how 'advanced' the techniques you use but rather on your ability to explain and develop a scalable machine learning approach to answering a real question.

Deliverable

You will have 4 weeks to work with your team to develop your analysis. Your submission summarizing and presenting your work will be a Jupyter notebook with the following 5 sections:

1. **Question Formulation** -- Introduce the goal of your analysis. What questions will you seek to answer, why do people perform this kind of analysis on this kind of data? Preview what level of performance your model would need to achieve to be practically useful.
2. **Algorithm Explanation** -- Create your own toy example that matches the dataset provided and use this toy example to explain the math behind the algorithm that you will perform.
3. **EDA & Discussion of Challenges** -- Determine 2-3 relevant EDA tasks that will help you make decisions about how you implement the algorithm to be scalable. Discuss any challenges that you anticipate based on the EDA you perform.
4. **Algorithm Implementation** -- develop a 'homegrown' implementation of the algorithm, apply it to the training dataset and evaluate your results on the test set.
5. **Application of Course Concepts** -- Pick 3-5 key course concepts and discuss how your work on this assignment illustrates an understanding of these concepts.

Think of this notebook as a communication mechanism or summary report. Please do not include scratchwork. Please *do* cite your sources in all 5 sections including the data source itself. Below we've included a rubric describing how your notebook will be evaluated.

Dataset:

The data for this project are available here:
http://labs.criteo.com/2014/09/kaggle-contest-dataset-now-available-academic-use/

Read more about the data at the Kaggle competition website here:
https://www.kaggle.com/c/criteo-display-ad-challenge

Resources:

Note that 'Click Through Rate Prediction' is not a single algorithm like 'Naive Bayes' but rather a goal which can be achieved through a number of different methods. There is a lot of literature out there about binary classification, ensemble methods, factorization machines, collaborative filtering and about the original Kaggle Competition. Do not feel pressured to implement any one approach -- instead try to get a sense for the space and then quickly narrow down an approach you will wrap your head around. Here are some reading materials to get you started.

- https://www.dropbox.com/s/s4x7wp8gjsh021d/TISTRespPredAds-Chappelle-CTR-Prediction-2014Paper.pdf?dl+=0
- https://www.csie.ntu.edu.tw/~r01922136/slides/ffm.pdf
- https://www.dropbox.com/s/iozods194twg2pv/MLParis2015-excellent-Sldies.pdf?dl=0
- http://statweb.stanford.edu/~jhf/ftp/trebst.pdf
- https://www.dropbox.com/s/2n8uekjwpaur3bj/Deep-Learning-for-Criteo-Documentation.pdf?dl=0
- https://arxiv.org/pdf/1711.01377.pdf
- https://arxiv.org/pdf/1701.04099.pdf
- https://research.fb.com/publications/practical-lessons-from-predicting-clicks-on-ads-at-facebook/
- https://www.csie.ntu.edu.tw/~r01922136/slides/kaggle-avazu.pdf

Team & Submission

Your team should include 3-4 people who share your live session instructor (i.e. Saturday folks work with other Saturday folks). You are free to form your own team. Anyone who does not have a team will be assigned one by us in week 11. To submit your work please designate a single person on the team who will push the final report to their GitHub repo where we will grade it.

Course Concepts

Here is a starter list of course concepts you may wish to discuss in the last section of your project -- note that this list is not intended to be comprehensive, just a way to trigger your memory.

- scalability / time complexity / I/O vs Memory
- functional programming / higher order functions / map reduce paradigm
- bias variance tradeoff / model complexity / regularization
- associative/commutative operations
- race conditions / barrier synchronization / statelessness
- the shuffle / combiners / local aggregation
- order inversion pattern / composite keys
- total order sort / custom partitioning
- broadcasting / caching / DAGs / lazy evaluation
- GD - convex optimization / Batch vs stochastic
- sparse representation (pairs vs stripes)
- preserving the graph structure / additional payloads
- One Hot Encoding / vector embeddings / feature selection

- normalization
- assumptions (for different algorithms - for example OLS vs Trees)
- implications of iterative algorithms in Hadoop M/R vs Spark

Rubric:

Section	5 points	4 points	3 points	2 points
Notebook Presentation	Report reads as a clear exposition of a thought process in 5 stages. Typo and error free. Well commented code + written discussion used to present the work. Cites sources without quoting for quoting's sake.	Notebook organization is clear. All 5 sections are clearly identified. Written descriptions introduce code blocks and clearly tie together a narrative.	Notebook organization is clear, but section content may not match instructions. May include scratchwork or big uncommented code blocks. Lacking narrative flow.	Notebook organization is unclear. Missing team names, course name, date and/or citations.
Question Formulation	Data set contents and context are clearly introduced. Clearly articulated question that is appropriate to the both the data and the algorithm and takes limitations of the data and/or algorithm into account.	Data set contents and context are clearly introduced. Report clearly forms a question that could be answered by this data and this algorithm.	Data set is clearly described in terms of its contents. Some attempt is made to discuss how the algorithm could apply to this dataset.	Data set is introduced without clear framing of the question that will be tackled.
Algorithm Theory	Math clearly explained without overly technical language Toy example is appropriate to the algorithm & dataset.	Some attempt to explain the math but explanation may be vague or overly jargony. Toy example matches the data and calculations	Some attempt is made to explain and demonstrate the algorithm. Toy example may not be well defined or may be ill suited to the	Theory is explained without reference to specific calculations or a specific toy example.

	Toy example calculations are clearly presented	are present but may not be well demonstrated.	data/ problem at hand.	
EDA	EDA tasks are well chosen and well explained. Code is scalable & well commented. Written discussion connects the EDA to the algorithm and/or potential challenges.	EDA tasks are clearly introduced both in terms of what the code does and why it is of interest. Code is clearly commented.	2-3 EDA tasks are performed but insufficient context is provided as to why these EDA tasks matter and/or what we learn from them.	The EDA performed is not relevant to the task at hand or only relevant superficially.
Implementation	Implementation matches theory discussed in section 2 and is scalable. Design choices are explicitly discussed using both text and code comments. Results are both clearly presented and clearly discussed	Implementation matches the Algorithm theory in section 2. Design choices are explicitly discussed. Results are presented but may be unclear or missing discussion.	Implementation may not match algorithm discussed in section 2 or there is not enough explanation of the design to tell.	Implementation is not scalable and or no attempt made to explain the design.
Course Concepts	Correctly identifies 3-5 course concepts that are relevant to this algorithm. Discussion demonstrates understanding of the chosen concepts *and addresses how*	Correctly identifies 3-5 course concepts that are relevant to this algorithm. Discussion demonstrates understanding of the chosen concepts.	3-5 concepts are clearly described but may not be relevant to the analysis performed or may not be related to the course.	Conceptual discussion demonstrates misunderstanding of one or more course concept

	the algorithm performance/scalability/ practicality would have been different given a different implementation.			

Total = ___ / 30