BDCC 19/20
Worksheet #3
May 7th, 2020

Questions about paper:

- **Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence**

1. Figure 1 shows the standar Python ecosystem for machine learning, data science, and scientific computing. Search the internet and write a paragraph about each of the tools appearing in the blue boxes. What is the tool in the orange box? What is it for and how does it work?

Pandas is a Python library for data analysis. Started by Wes McKinney in 2008 out of a need for a powerful and flexible quantitative analysis tool. It implements almost all code for data analysis, statistical analysis, data I/O, and integrates with databases and other software such as SPSS and SAS. It also integrates with the google cloud platform (GCP).

Scikit-Learn is a machine learning library that implements most of the supervised and unsupervised methods plus validation, and metrics.

Network-XGraph Analytics implements methods and fucntions to handle graph networks.

PyTorchChainerMxNet Deep Learning implements methods and functions for deep learning

pyViz Visualization for data visualization is a collection of functions ans methods based on other python modules that offer all kinds of data visualization, including maps.

The orange box shows Dask, a module in python that helps to parallelize and stream data.

2. What are the differences between Spark and Dask?

Dask implements most of the spark concepts, but it is lighter and smaller. It also couples well with python numerical libs like numpy. Dask is python code. Spark is written in Scala and can be used with several other languages, including python.

3. “Pandas’ data frame format uses columns to separate the different fields in a dataset and allows each column to have a different data type...Laying out the data contiguously by column enables SIMD by allowing the processor to

group, or coalesce, memory accesses for row-level processing,making efficient use of caching while lowering the number of accesses to main memory.”
Given the sentence above, what is the consequence of this memory arrangement in an algorithm that will cluster instances using, for example, a k-means algorithm, where distances between pairs of rows need to be calculated?

This arrangament of data in memory will not work very well if the data processing needs to be done per row instead of column. Processing will be slow due to jumps in memory.

4. Figure 2 shows a python code example. What does the `make_pipeline` function do?

It creates a sequence of estimators to be executed. Similar to scikit-learn Pipeline. Not to be confused with beam.Pipeline.

5. Why is bagging easier to parallelize than boosting?

Bagging is an ensemble method where the generation of each model is indenpendent on the other. On the other hand, each boosting step depends on the previous one, since the weights of misclassified examples are adjusted according to weights assigned in previous runs.

6. What is the purpose of Neural Architecture Search? (Section 3.3)

To “learn” the best neural network strcuture among various structures by optimizing various parameters automatically.

7. Explain the sentence: “The CoCoA framework preserves locality across compute resources on each physical machine to reduce the amount of network communication needed across machines in the cluster”

Distributed programming can incur a high overhead if message passing happens very often. In order to reduce the number of messages being exchanged between machines, CoCoa defines meaningful subproblems for each machine to solve in parallel, and then combines updates from thesubproblems in an efficient manner.

8. What is the relation between Dask, RAPIDS and cuML?

Dask is a python library meant to parallelization and streaming. RAPIDS is almost a replacement of Pandas, Scikit-learn, and Network-X for GPUs, named cuDF, cuML and cuGraph.

Questions about paper:

- From Persistent Identifiers to Digital Objects to Make Data Science More Efficient

1. Given the sentence below, what is necessary to make data of better quality and better organized?

“The biggest factors for these inefficiencies seem to be a bad quality of data/metadata and a bad data organization, i.e., if a rough description of data is available, it is difficult to find out how to access them, where the corresponding metadata is to enable processing, etc.”

It is necessary to define a common standard that can be understood by anyone in a non-ambiguous way.

2. Do you think that the points mentioned in the paper are the only ones to make data of better quality? (you may need to complement your reading with this report).

The definition of a standard may improve data quality. However, data quality will always be dependent on who inputs the data. If there is no input checking against the standard or some checking was not considered during the design of the standard, data may still become low quality.

3. What is a Digital Object and how can it help to improve data access?

A DO is a Persistent Identifier (PID) that can represent a single object or a collection. Like an URL, but used to identify data. It can help improving access as long as the same digital object will refer always to the same data.