CSE545 Sp20 - Assignment 2 Description

Big Data Analytics
Stony Brook University
CSE545 - Spring 2020

Assignment 2.
Assigned: 3/7/2020; Due: ~~3/26/2020~~ 4/4/2020 11:59pm

Overview.

Part I Goals.

- Study a single big data library or tool, in-depth
- Practice summarizing a potential complex topic into usable information, distilling it down to the important points.
- Build a guide that helps yourself and your classmates in determining which modern big data libraries and tools are available for their project goals.
- Practice information investigation with a partner.

Part II Goals.

- Attempt analytic problem solving from Spark and Tensorflow perspectives
- Practice working with spark transformations
- Practice working with a tensorflow graph
- Review auto-diff
- Explore Minhashing and Locality Sensitive Hashing

Submission. Submission will take place in two pieces (each worth 50%):

Part I (2 person team): The big data toolkit guide must be inserted as a page into the shared google doc by the deadline.

Part II (Individual Assignment): Must be placed in a single pdf document, with one question answer per page, clearly label. The pdf must have machine readable text (i.e. one can highlight and copy-paste the text; images may be used for figures but not for text). Submit the pdf to blackboard. Please only submit a pdf file to blackboard.

Academic Honesty. Copying chunks of code or problem solving answers from other students, online or other resources is prohibited. You are responsible for both (1) not copying others work, and (2) making sure your work is not accessible to others. Assignments will be extensively checked for copying of others' work. Problem solving solutions are expected to be original using concepts discussed in the book, class, or supplemental materials but not using any direct code or answers. Please see the syllabus for additional policies.

Part I: Big Data Toolkit Guide (2 Person Teams).

Here, you along with your partner will do your part in a class-wide collaboration to produce an invaluable tool for yourself and your classmates, **The 2020 BDTK Guide** (The 2020 Big Data ToolKit Guide). The BDTK will be a document describing recent key libraries and tools for solving big data problems.

Motivation for BDTK Guide. Modern big data analytics stands on the shoulders of giants: we could not achieve what we hope if we started each project from scratch. Organizations and developers working in big data realize this and produce a steady stream of new tools to address the common challenges to analyze wide and tall data. At this point, at least hundreds of such tools and toolkits are available, but which ones will be useful to you or your future projects?

The main focus of CSE545, the class, is to give you fundamental knowledge of big data such that you can tackle a variety of situations yourself, but you shouldn't always need to reinvent the wheel from the basics when others have been perfecting the wheel you need potentially for years or decades. Thus, each team (2 people) will investigate a single topic.

Choosing a topic.
You must pick from a list of pre-approved topics or you may suggest your own and seek approval for it by messaging the "instructors" via Piazza. Sign up here by 3/11.

Criteria for your Contribution.
You will be given 1 page in which you must cover:

- **Introduction:** (*5 pts*)
  Why should we care about this technology? How is it related to Big Data?

- **Approach:** (*5 pts*)
  How does it work? Explain the algorithm or framework.
- **Results:** (*5 pts*)
  Are there benchmarks for its use? How does it compare to similar technology?
- **Pro-Cons:** (*5 pts*)
  What are the good aspects, what are the bad aspects? Be sure to add a sentence on "**contributor thoughts:**" What are your own unique thoughts on the pros and cons of the technology? Do you envision an extension that might be helpful?
- **Conclusion:** (*5 pts*)
  Summarize the 2 to 4 points you think are most important

Concise, information rich content. For each of the sections above you will not simply be graded on having content but on the quality of the content and how well it answers the questions in concise, clear, and engaging terms. (*15 pts*)

Referenced Content. Referencing information can be a strong way to support your contributions, but each section of contribution above needs to still be a "contribution" rather than just be a pasting or simple paraphrase of another website. If you quote content from another page, you must use quotes and cite it. If you paraphrase a statement from others or a website, you must cite them. If you take or build on someone else's figure then you must cite them. It is fine, ethically, to do so as long as cited, but if you mostly just quote others and take their figures, then the lack of novelty will be reflected in your grade as less of a contribution.

Style.
In order to make our guide consistent and visually appealing, as well as to make evaluation of your work similar, each page was conform to the following specifications:

- Margins: approx. 0.5" on all 4 sides.
- Columns: 2 with approx. 0.3in margin; justified text
- Fonts:
  - Body text: Times New Roman, 11pt.
  - Section headings: Calibri 13pt bold-Italic
  - Within captions, tables, figures, or images: Calibri 9 - 11pt.
- Line Spacing:
  - Body text: Single (1.0)
  - Section headings: 6pt spacing above heading

The total number of tables plus figures should be 2 or 3.
(*10 pts*)

An example from the TAs is here.

Let's make a lasting, easy to reference, document for each other. Again, do not expect a good grade if all you do is fill up a page. It must be composed of useful information that your team put together itself, building on information you find by researching the tool.

Part II: Problem Solving (Individual Assignment)

Here, you, by yourself, will review course material in order to provide an answer and justification to each question below. Turn in your solution within a single, machine readable pdf (all text must be able to be highlighted -- no images of text; images of figures are ok). Place your answer to each question on a new page and clearly label the question number at the top: “Part II, Question 1”.

1. (*5 pts*) Adapt the MapReduce single pass matrix multiplication algorithm from MMDSv3 2.3.10 to run on Spark instead of tensorflow. To answer, use pseudocode along with a description of the purpose of each Spark transformation. Assume records are of the form: (matrix_name, ((i, j), v) where matrix_name is “M” or “N”, i and j correspond to row and column respectively, and v is the value. State all other assumptions clearly.

2. (*10 pts*) In class, we went over tensorflow code to define the linear regression model and perform gradient descent over its beta parameters: demo code here.

   In such a case, during each iteration, the parameters (betas) are updated by subtracting their gradients times the learning_rate (alpha). One can think of the gradients as a speed where each iteration’s change depends only on the current gradient. An alternative approach, *momentum optimization*, keeps track of the previous speed (i.e. a momentum) and then treats the gradient more like an acceleration or deceleration, telling the descent whether to slow down or speed up in each dimension. This is discussed in the HOML (Geron, 2017) book on pages 294 - 295 (distributed on Piazza under Educational Fair Use) -- note that the notation is different in this book than the other: θ represents the parameter vector (“betas” in our code), η represents the learning rate (“learning_rate” in our code), and β is a new term representing the momentum, or how much to pay attention to the previous speed (represented by **m**).

   Rewrite our linear regression gradient descent demo algorithm to instead use momentum optimization. **You must specify the new training operation yourself.** You may **not** use any tf.train Optimizer. Assume the variable X is any 2d tensor of features and y is a 1d tensor of corresponding outcomes for each row of X. State all other assumptions clearly.

3. (*20 pts*) Consider the minhashing technique we went over in class.
(a) Prove that if the Jaccard similarity of two sets is 0 then the estimate one would get from minhashing always yields 0.

(b) Consider that a massive characteristic matrix is stored in HDFS such that each record is a set of elements for a given document:

    (set_name, (elem1, elem2, ...))

Where only the elements present in the document are stored in the list (i.e. this is essentially a sparse representation of the characteristic matrix; it doesn’t need to store the zeros). These are further divided into chunks automatically across hdfs and the entire characteristic matrix is way too large to keep in memory on a single machine. Design a Spark-based solution to produce a signature matrix using efficient minhashing. The resulting characteristic matrix should have only 500 rows, no matter the size of the original characteristic matrix. You may use a combination of pseudocode and natural language to describe your solution. Justify all steps and clearly state your assumptions (hint there are multiple effective approaches to this).

4. (*15 pts*) Consider running locality sensitive hashing on the resulting characteristic matrix produced from your solution to 3(b).
(a) If you used a band size of 5 and a particular pair of documents (set1, set2) had a Jaccard Similarity of 0.75, what would be the probability that the two sets matched in at least 1 band.

(b) Suppose you wanted LSH to capture nearly all candidate pairs, 99% of all possible pairs, with at least 90% Jaccard similarity. At the same time, you want to keep the false-positive rate fairly low (e.g. below 25%). How can you select the number of bands optimally? Describe a solution and justify why you expect it to result in a set of candidate pairs such that 99% of all sets that match at 90% Jaccard Sim will be captured and yet only 25% of the resulting candidate pairs will have < 90% Jaccard similarity.

(c) Suppose you wanted to speed up your approach to LSH in 3(b) by allowing a greater number of false positives than 25%. Describe how you would adjust your approach and justify how it would speed up the solution. State your assumptions.