Homework 3

17-400/17-700: Data Science and Machine Learning at Scale

Due Friday, October 16th at 11:59 PM

1 Introduction

This assignment builds upon the Spark and ML knowledge you've gained in the previous homeworks to further flesh our your distributed machine learning expertise.

This assignment consists of two major sections. You will first go through the steps for creating a click-through rate (CTR) prediction pipeline. You will then implement some popular gradient descent algorithms in a distributed-friendly manner.

2 Logistics

As with the previous homeworks, we provide the code template for this assignment in a Jupyter notebook. What you need to do is to follow the instructions in the notebook and implement the missing parts marked with `'<FILL IN>'` or `' YOUR CODE HERE'`. Most of the `'<FILL IN>/YOUR CODE HERE'` sections can be implemented in just one or two lines of code.

In addition, in this homework we ask you to fill out a written answer at the end of the second section which we will manually evaluate for grading. There is also the opportunity for extra credit in two parts which we will also assess manually.

2.1 Getting lab files

The code for this homework is in a single notebook file. You can obtain the notebook `'assignment_notebook.ipynb'` after downloading and unzipping `hw3.zip` at
https://github.com/17-700/released-hws-fa2020/raw/master/hw3/hw3.zip.

Next, as for Homeworks 1 and 2, import the notebook into your Databricks workspace. You can refer to the instructions of the previous homeworks if you need a refresher on how to set up your environment - the requirements are identical for this one.

2.2 Preparing for submission

We provide several public tests via `assert` in the notebook. You may want to pass all those tests before submitting your homework. Also be sure to fill out the manually graded questions in the cells that we provide you.

In order to enable auto-grading, please do not change any function signatures (e.g., function name, parameters, etc) or delete any cells. If you do delete any of the provided cells (even if you re-add them), the autograder will fail to grade your homework. If you do this, you will need to re-download the empty 'assignment_notebook.ipynb' file and fill in your answers again and resubmit.

Also be sure to comment out any Databricks-specific functions such as `dbutils` before submitting your notebook. The autograder environment runs outside of Databricks, and so these function calls will fail and cause subsequent statements in the same cell to fail. This in turn might lead to cascading errors in later cells and unintuitive errors in the autograder output.

2.3 Submission

1. Export your solution notebook as a IPython notebook file on Databricks via `File -> Export -> IPython Notebook`
2. Submit your solution via Gradescope (Please don't rename your notebook file).

3 Section I: Click-Through Rate Prediction

In this section, you will go through the steps for creating a click-through rate (CTR) prediction pipeline. You will work with the Criteo Labs dataset.

This section covers:

- Featurizing categorical data using one-hot-encoding (OHE)
- Constructing an OHE dictionary
- Parsing CTR data and generating OHE features
- CTR prediction and logloss evaluation
- Reducing feature dimensionality via feature hashing

4 Section II: Gradient Descent Optimization

In this section, you will build a distributed version of minibatch SGD along with two other stochastic gradient descent optimization algorithms. In particular you will be implementing:

- Minibatch SGD
- Adagrad
- Adaptive Moment Estimation (Adam)

This section also contains a manually graded question at the end along with an opportunity for extra credit.

See the notebook for detailed descriptions and instructions of each question.