Homework 4 Part B

10-405/10-605: Machine Learning with Large Datasets

Due Wednesday, April 21st at 11:59:59 PM Eastern Time

Instructions: Submit your solutions via Gradescope, following the template below. Note that this assignment is the second part of HW4; similarly to HW4A it does not contain a theoretical written part and the programming part is much more open-ended than before. Because of this, we will not be autograding your submission. Instead, you will submit a report consisting of responses to questions asked in the notebook. The report is worth 100% of your grade. We may refer to your code submission in case we have doubts. **You must still submit your code to Gradescope under the HW4B Programming submission slot in order to be given credit for the written.**

Submitting via Gradescope: When submitting on Gradescope, you must assign pages to each question correctly (it prompts you to do this after submitting your work). This significantly streamlines the grading process for the course staff. Failure to do this may result in a score of 0 for any questions that you didn't correctly assign pages to. It is also your responsibility to make sure that your scan/submission is legible so that we can grade it.

1 HW4 Part B: Machine Learning with the Million Song Dataset

1.1 Introduction

The goal of this assignment is to gain hands-on experience with training a machine learning model on a large dataset on cloud computing services.

You have already set up your AWS EMR and run Spark in the jupyter notebook in **Part A**. In **Part B**, you will do EDA, data cleaning, and modeling in the jupyter notebook we provided with instructions and some (very limited) starter code. At the end, you will have a chance to shine by optimizing your pipeline and model.

1.2 Logistics

We provide the code template for this assignment in *one* Jupyter notebook. **We also provide the converted MSD in a public S3 bucket**. Although you may have produced the exact same `csv` files in **Part A**, you are required to use our version of the dataset to avoid unexpected inconsistency. Failure to do so may result in point deductions.

You should follow the instructions in the python notebook and implement the missing parts marked with '`# YOUR CODE HERE`'. **Unlike the previous homework, this homework gives you more freedom and less guidance so you may need to complete a cell from scratch.** That said, each cell will be a small task which will be easy to implement following our instructions.

Note that we will not autograde your code through Gradescope. Instead, you should submit your code along with your report (including plots, statistics, and short answers). Points will be given according to your answers in the report. We may still refer to your code submission in case we have doubts.

1.3 Getting lab files

You can obtain the start code for **Part B** `hw4-b.ipynb` after downloading and unzipping `hw4-b.zip` from https://github.com/10605/released-hws/releases/tag/s21-hw4b.

1.4 Preparing for submission

Complete the `hw4-b.ipynb`. Complete corresponding written questions using this write-up `hw4-b.pdf` as a template.

1.5 Submission

1. Download the notebook to your local computer by going to `File -> Download as -> Notebook (.ipynb)` and complete relevant written questions in this document.

2. You only need to submit `hw4b.ipynb` and this write-up `hw4b.pdf` to Gradescope.

2 Part B: Modeling

In this part of the homework, you will perform EDA and data cleaning, and then train some models with the original features. You will then do some feature engineering similar to what we did in homework 1 (TF-IDF and BoW), and then train models with these new features.

IMPORTANT: to make grading consistent and to make those who failed to finish Part A still able to work on the rest of this homework, you are asked to switch to our S3 bucket for the converted MSD dataset. The bucket name is already hard-coded in the starter code in the notebook so you don't have to do anything. Just make sure you don't change it.

2.1 Exploratory Data Analysis

(a) *[2 points]* Explain why the two features seem problematic (after performing .summary() operation).

(b) *[3 points]* Histograms (remember to label them)

(c) *[3 points]* Explain what is weird about `year`'s distribution and what might cause this. Describe how you could filter `year` to make its histogram look more balanced.

(d) *[2 points]* New histogram for `year`.

(e) *[5 points]* Plots for the three pairs. Describe your findings.

(f) *[3 points]* Think about what simple technique you could use to visualize large datasets while retaining data distribution. Briefly describe what you did.

2.2 Data Cleaning

(a) *[2 points]* Your justification for dropping the two features.

(b) *[5 points]* Compare the two numbers and explain the advantages and potential problem of doing this step. What other techniques could you use to potentially do better?

(c) *[2 points]* State the two features.

(d) *[6 points]* Explanation your proposed solution and discuss its pros and cons.

(e) *[2 points]* Report the percentage:

2.3 Baseline

(a) *[3 points]* Explain why treating this as a classification problem might be a sensible choice.

(b) *[2 points]* Report what percentage of songs are assigned the “popular” label.

(c) *[2 points]* Explain why we shift the year.

(d) *[5 points]* Explain what scaling means and why we want to perform scaling before the learning step.

(e) *[5 points]* Explain the difference between these two metrics and when AUC might be more useful than accuracy.

(f) *[8 points]* Calculate the train and test AUC of both models and report them.

Models	Train AUC	Test AUC
Logistic Regression		
Random Forest		

2.4 Featurization: Bag-of-Words and TF-IDF

(a) *[3 points]* Explain what the `vocabSize` hyperparameter means in the context of Bag-of-Words.

(b) *[3 points]* Other than featurizing texts, what other feature engineering would you do on the dataset? Briefly describe one.

(c) *[3 points]* Explain where this number "31" comes from.

2.5 Modeling with New Features

(a) *[8 points]* Evaluate train and test AUC for each model and report them.

Models	Train AUC	Test AUC
Logistic Regression		
Random Forest		

(b) *[8 points]* Include the plot and your explanations.

2.6 Do Your Best

(a) *[2 points]* Your final AUC:.

(b) *[4 points]* Your model and hyperparameters.

(c) *[4 points]* Describe your approach.

2.7 Reflection

[5 points] What challenges did you face in HW4 Part B? How did you overcome these challenges? What did you learn from HW4 (both Part A and B)?”

3 Collaboration Questions

1. (a) Did you receive any help whatsoever from anyone in solving this assignment?

   (b) If you answered ‘yes’, give full details (e.g. “Jane Doe explained to me what is asked in Question 3.4”)

2. (a) Did you give any help whatsoever to anyone in solving this assignment?

   (b) If you answered ‘yes’, give full details (e.g. “I pointed Joe Smith to section 2.3 since he didn’t know how to proceed with Question 2”)

3. (a) Did you find or come across code that implements any part of this assignment?

   (b) If you answered ‘yes’, give full details (book & page, URL & location within the page, etc.).