Scalable Data Mining (Autumn 2023)

Assignment 2 (Total Marks: 100)

Part A: Spark (Full Marks: 50)

Steps for Spark installation:

1. Follow the guidelines given in this link to install Spark in your system: https://medium.com/@josemarcialportilla/installing-scala-and-spark-on-ubuntu-5665ee4b62b1

Instructions: Please submit your answers (code+output+your way to approach the problem) to the following questions as a write-up in a PDF file via Moodle.

Question 1 (Marks = 10)

In this assignment, you have to use Spark to have a look at the Movie Lens dataset containing user-generated ratings for movies. The dataset comes in 3 files:

- ratings.dat contains the ratings in the following format: UserID::MovieID::Rating::Timestamp
- users.dat contains demographic information about the users: UserID::Gender::Age::Occupation::Zip-code.
- movies.dat contains meta-information about the movies: MovieID::Title::Genres

Please read the readme file in the zip folder for further information.

==========

(10 points):

a) Download the rating file, parse it, and load it in an RDD named ratings.

b) How many lines do the ratings RDD contain?

Question 2 (Marks = 40)

Using the same data file from Question 1, perform the following operations:

(20 points):

a) Read the movies and user files into RDDs. How many records are there in each RDD?

b) How many of the movies are a comedy?

(20 points):

c) Which comedy has the most ratings? Return the title and the number of rankings. Answer this question by joining two datasets.

d) Compute the number of unique users that rated the movies with movie_IDs 2858, 356, and 2329 **without using an inverted index**. Measure the time (in seconds) it takes to make this computation.

Submission Instructions for Spark:

You will submit 1 file using the filename *RollNo_AssignmentNo_Spark.pdf* with the following details:
(1) description/logic of how you are going to use Spark to solve each problem using Scala,
(2) the code snippets for each problem and
(3) their respective outputs.

Part-B: Pytorch (Full Marks: 50)

Question 1: (50 marks)

Introduction:
In this assignment, you will delve into the world of activation functions by implementing and comparing two innovative options: **Swish** and **Gelu**. Activation functions play a crucial role in the performance of neural networks, and this assignment will give you a hands-on experience in working with these functions to assess their impact on a real-world dataset. This link describes the Swish and GeLu activation functions:
https://www.v7labs.com/blog/neural-networks-activation-functions

Aim:
The aim of this assignment is to implement custom activation functions in Pytorch, the Swish and Gelu activation functions, and compare their performance with the previously used ReLU and Sigmoid functions.

Dataset Selection:
Utilize the **normalized Boston Housing Dataset** that was employed in Assignment 1.

Network Reuse:
Reuse the neural network architecture that you implemented in Assignment 1. This network will serve as the foundation for evaluating the Swish and Gelu activations.

Activation Function Implementation:
Implement the Swish and Gelu activation functions for both forward and backward propagation in your neural network by extending the torch.autograd.Function function. You will need to modify the network code to incorporate these new activation functions.

Activation Function Usage:
Replace the previously used ReLU and Sigmoid activation functions in your neural network with the newly implemented Swish and Gelu functions. Ensure that the network is properly updated to accommodate these changes.

Performance Comparison:

1. Train and evaluate your neural network with both Swish and Gelu activations separately.
2. Use the training combinations which gave you best results for first assignment.
3. Use appropriate evaluation metrics such as **Mean Squared Error (MSE)** to compare the performance of the network with Swish and Gelu activations against the original ReLU and Sigmoid activations.

Report:

1. Provide a detailed explanation of the implementation of Swish and Gelu activations in your network.
2. Present a comparative analysis of the model's performance with Swish and Gelu against ReLU and Sigmoid. Discuss any observed differences in training and convergence.
3. Summarize your findings and draw conclusions regarding the suitability of Swish and Gelu activations for regression tasks on the Boston Housing Dataset.

Submission Details for Pytorch:

You should submit the following for both Spark and Pytorch in zipped format (Rollnumber_AssignmentNo_Pytorch.zip):

- **Report** with all the contents as mentioned under the **'Report'**. Analyze the observations and explain them. **(30 marks)**
- **Python codes:** You can add functions as per requirements. **(20 marks)**

Overall Submission Details:

Please make one folder named Rollnumber_AssignmentNo containing two folders one for Spark and one for Pytorch and then zip it and upload it in Moodle.