2024 AIML427 Big Data: Assignment 3

This assignment has 100 raw marks and is due on ***11:59 pm, Tuesday, 11 June 2024***. Please submit your report as a single *.pdf* file. Make sure you read the *Assessment* and *Submission* sections at the end of this file. This assignment contributes 30% to your overall course grade.

Objectives

Spark MLlib and Spark ML are libraries that support scalable machine learning and data mining algorithms such as classification, clustering, collaborative filtering and frequent pattern mining. This assignment is on how to use these libraries to solve a given problem.

Another goal of this assignment is to improve your teamwork skills such as organising, planning, communication, and collaboration. The group work part of this assignment is designed as a teamwork of **a maximum of 3 people**. Each group is encouraged to have members from different backgrounds so that each member can learn from each other and cooperate to finish the work.

This assignment also aims at improving students' presentation skills. You need to **present your individual project in *4-5 minutes***. *Note that late days cannot be applied to the presentation portion.* The presentation slides of each student (PowerPoint file is preferred) have to be submitted to the system by ***11:59 pm, Tuesday, 11 June 2024***. Presentation date and time slots will be given later.

Question Description

1) Group work (40 marks): Implement two programs that apply Spark's Decision Tree algorithm and Logistic Regression algorithm to the provided KDD dataset. Run these programs 10 times on the school's Hadoop cluster, each using a different seed to split the dataset into a training and a test set. Write *less than 6 pages* to:

   (a) (12 marks) Describe the two programs using pseudo-code;

   (b) (6 marks) Describe how to install and run the two programs step by step in a Readme.txt file.

   (c) (10 marks) Report the training and test results including the max, min, average accuracy and the standard deviation obtained from the 10 runs, and the running time of each program.

(d) (12 marks) Compare and discuss the results of the two models.

The KDD dataset contains a standard set of data which includes a wide variety of intrusions simulated in a military network environment. It has 2 classes, normal connection and anomaly connection (ie. attack), and 41 features (duration, protocol_type, service, flag, src_bytes, dst_bytes, land, wrong_fragment, urgent, hot, num_failed_logins, logged_in, num_compromised, root_shell, su_attempted, num_root, num_file_creations, num_shells, num_access_files, num_outbound_cmds, is_host_login, is_guest_login, count, srv_count, serror_rate, srv_serror_rate, rerror_rate, srv_rerror_rate, same_srv_rate, diff_srv_rate, srv_diff_host_rate, dst_host_count, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate, dst_host_srv_serror_rate, dst_host_rerror_rate, dst_host_srv_rerror_rate). You can get more information about this dataset here.

2) **Individual project** (60 marks): Solve a machine learning task with Apache Spark machine learning library. Chooses one machine learning task and a particular problem to solve, for example, predicting flight prices, grouping customers based on their behaviours, etc. The chosen dataset needs to be larger than $20MB$ and have at least 30 features after initial preprocessing.

   You are required to write a detailed report *in maximum 12 pages* and give a *4-5mins* presentation to:

   (a) (10 marks) Describe the task including the details of the input data (data source, data size, the original number of features and instances, feature types, whether missing values exist, etc.) and the expected output of the system;

   (b) (6 marks) Describe all the preprocessing steps applied to the download data file(s) to obtain the dataset that is used as input of your program, which must be at least $20MB$ and have at least 30 features.

   (c) (15 marks) Describe the program using UML class diagrams and/or pseudo-code;

   (d) (5 marks) Describe how to install and run the program step by step in a Readme.txt file. *(This can be skipped in the presentation.)*

   (e) (12 marks) Compare and discuss the results (including the training and test accuracy, the running time, the model, etc. depending on the program) of the program with and without normalising/scaling data.

   (f) (12 marks) Compare and discuss the results (including the training and test accuracy, the running time, the model, etc. depending on the program) of the program with and without transforming data using PCA.

   ***Note**: If the obtained models are too big, you can put them in the appendix.*

Assessment

You can use any font to write the report, with *a minimum of single spacing and 11 point size* (handwriting is not preferred).

Reports exceeding the maximum words or page limit will be penalized. Any additional material can be included in an appendix, which will not count towards the page limit. The final report should be submitted through turn-it-in as well. Please note the plagiarism score; more than 30% is problematic. The report should be your own individual work and late reports will be penalized.

The usual mark checking procedures in place for all assessment apply to this report. The assessment of the reports will account for the understanding of big data, clarity and accuracy of the answer, presentation, organisation, layout and reference.

In addition, one of the skills required of a scientist is the ability to communicate effectively. No matter what the scientific merit of a report, if it is illegible, ungrammatical, misspell, mispunctuated or ambiguous, then it is less useful and marks will be deducted.

Submission

You are required to submit *soft copy* including **the code, the dataset, the report, and the Readme files** of the required answers.

The soft copy of the document should be submitted through online submission system from the AIML427 course website *by the due time*. If the dataset is too large to submit, please put it in your cloud storage and put a share link in your report. There will be *two submission directories* for the group report and the individual report, respectively. Each group needs to select *a representative* who will submit the group report. Please submit *only one group report for one group*. The group report must list all group members clearly.