University of Aberdeen

School of Natural and Computing Sciences

Department of Computing Science

MSc in Artificial Intelligence

2020 - 2021

Please read all the information below carefully

Assessment Item 2 of 2 Briefing Document – Individually Assessed (no teamwork)

CS551G - Data Mining and Visualisation

Note: This assessment accounts for 50% of your total mark of the course.

Learning Outcomes

On successful completion of this component a student will have demonstrated competence in the following areas:

- Using a non-trivial dataset, plan, execute and evaluate significant experimental investigations using multiple data mining, visualisation and machine learning strategies

Information for Plagiarism and Conduct: Your submitted report and source code may be submitted for plagiarism check (e.g., Turnitin). Please refer to the slides available at MyAberdeen for more information about avoiding plagiarism before you start working on the assessment. Please also read the following information provided by the university: https://www.abdn.ac.uk/sls/online-resources/avoiding-plagiarism/

In addition, please familiarise yourselves with the following document "code of practice on student discipline (Academic)": https://tinyurl.com/y92xgkq6.

Application Problem Definition: Gas Turbine Availability (Anomaly Detection)

The objective of this assessment is to analyse a large dataset concerning gas turbine (figure 1) data, specifically on the operational data involved, such as temperatures, flow rates, pressures, and vibrations. Since their introduction in the late 1930's gas turbines have often been found to be critical pieces of infrastructure for their owners who have invested considerable sums in their installation, maintenance, and operation. Whilst expensive they have some unique features compared to other power sources such as low moving part count, high reliability and compact nature. Due to the integral and often critical nature of the role that gas turbines play in their operators businesses, it is often paramount that they have the turbines available as much as possible, or know well enough in advance that an issue is occurring that they can operationally deal with any such outage.

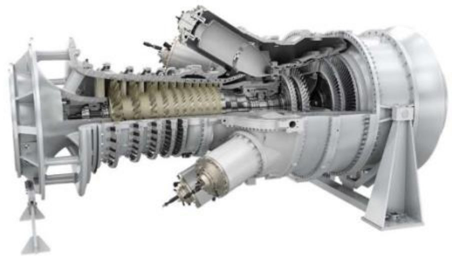

Figure 1. Gas Turbine @https://www.siemens-energy.com/global/en/offerings/power-generation/gas-turbines/sgt-400.html

****Please read all the information below carefully*****

The dataset can be downloaded from MyAberdeen. It is based on data from a research project that investigated how to detect anomalies and events in gas turbines. The dataset includes two classes (normal/abnormal condition) and a number of features, which will need to be utilised throughout this assessment. The class membership of each row is stored in the field 'Status'. The task is to develop a set of classification models for automatically classifying reactors as normal or abnormal, based on their parameters/features. No prior knowledge of the domain problem is needed or assumed to fulfil the requirements of this assessment.
Feature information in the dataset include:

- Temperatures across the gas path of the engine, bearings, and ambient conditions.
- Flow rates at various points
- Pressures at various points
- Vibrations at various points

Status refers to the condition of the gas turbine, or in other words, we consider this to be our label/annotation for the sake of all implementations (first column). As this is a binary classification task, all implementations should treat the problem as such.

Unit of measurement or range of values of each feature are not relevant. However, features can be at different scales and/or measured in different units.

Report Guidance & Requirements

Your report must conform to the below structure and include the required content as outlined in each section. Each subtask has its own marks allocated. You must supply a written report, along with the corresponding source code written in **python**, containing all distinct sections/subtasks that provide a full critical and reflective account of the processes undertaken.

This assessment revolves around big data mining processes and visualisation, which as you already know, does not necessarily imply that the datasets used should be huge, but rather that the focus is on distributed processing that can be scaled on demand in a fault tolerant way. Therefore, we will focus on the fundamentals of distributed data mining and the use of apache ecosystem. Apache ecosystem covers a number of different things needed in the realm of data mining and machine learning, including but not limited to distributed real-time computational systems, streaming dataflow engines, distributed analytics and machine learning platforms to name a few. The following tasks require you to expand and elaborate upon the principles of big data mining, different components of the apache ecosystem and some aspects on how such techniques can be used in real-life problems.

Task 1: Description of Distributed Learning Big Data Ecosystem (10/50) – (~max 600 words)

Using your own words, the lecture material and any other relevant sources, explain the distributed big data processing ecosystem. Your description should cover the following points at a technical level:

- Distributed File Systems (e.g. HDFS)
- Resource Manager and Scheduler (e.g. YARN)

Please read all the information below carefully

- Volume, Velocity, Variety, Value, Veracity (Big Data Characteristics)
- Fault Tolerance and Resilience
- Apache Spark
- Spark ML/PySpark
- SparkSQL
- Docker containers (e.g. Kubernetes)

Task 2: Develop distributed models in apache spark to classify gas turbines (40/50) – (~max 1800 words)

The problem we aim at tackling has been clearly described and defined earlier. This task includes ***five*** subtasks, each of which bears its own marks.

Subtasks:

1. Create an apache spark environment and load the dataset provided. You may use CoLab or a jupyter notebook. Please create a table providing summary statistics of this dataset, i.e. mean values, range, standard deviations, min/max values, median values and 25%/50%/75% percentile values. Comment on whether there are any missing values present throughout. The column that contains the labels (normal/abnormal) should not be part of the table (**3 marks**).
2. Visualise the data as follows: create two plots, i.e. box plot and a scatter plot. The box plot shall include the two classes (normal/abnormal) in the x-axis, and the "Vibration_Sensor_1" in the y-axis. The scatter plot shall include the feature "Vibration_Sensor_2" with the graphs of both classes appearing in the same plot (different colour or symbol). Please elaborate on what information one can obtain from each of these two plots (**hint: handyspark)** (**7 marks**).
3. Using the apache spark ecosystem, such as pyspark, sparksql or any other component needed, please develop and train a random forest model with a binary output to classify the condition of gas turbines, i.e. normal or abnormal, based on how they operate. Split the dataset provided into training (70%) and test (30%) sets. Please use the training set to train your developed model, keeping the test set only for evaluating its performance in unseen data (**15 marks**).
4. Use the following three metrics to report the model's performance, i.e. Precision/Recall, Accuracy and Area under the curve (AUROC). When reporting performance, please only use the test set created by yourselves (**5 marks**).
5. Repeat steps 3 and 4 but using a multilayer perceptron classifier this time. (**10 marks**)

****

Bonus – Optional: Should you decide to try an alternative task, there will be a bonus of **5 marks.** The maximum overall mark for this assessment remains at 50/50; however, attempting the bonus exercise will a) make you practice with an alternative distributed library and b) enhance your chances of getting a higher mark overall.

Description: Using the package elephas and getting inspiration from the colab notebook that has been provided to you, create a distributed keras model - (model = sequential ()) – with:

a) three dense layers,
b) relu activation function, and

Please read all the information below carefully

c) dropout layers after each dense layer.

Last layer will have a softmax activation. Please use categorical_cross entropy and adam as optimizer. Learning rate can be chosen by you, which can be used to finetune / optimise performance.

Report performance in terms of accuracy only (test set).

****

Marking Criteria

- Quality of the report, including structure, clarity, and brevity
- Reproducibility. How easy is it for another MSc AI student to repeat your work based on your report and code?
- Quality of your experiments, including design and result presentation (use of figures and tables for better reporting)
- Configured to complete the task and the parameter tuning process (if needed)
- In-depth analysis of the results generated, including critical evaluation, insights into data, and significant conclusions
- Quality of the source code, including the documentation of the code

Submission Instructions

You should submit a PDF version of your report including code snippets via MyAberdeen by **23:59 Sunday 28th March 2021**. The name of the PDF file should have the form "CS551G_Assessment2_< your Surname>_<your first name>_<Your Student ID>". For instance, "CS551G_Assessment2_Smith_John_4568985.pdf", where 4568985 is your student ID.

In addition to the written report, you should also submit supplementary material in the form of a zip file containing the source code of your implementation (ideally as a python notebook ".ipynb"). The naming convention should follow the same form as for the PDF. For instance, "CS551G_Assessment2_Smith_John_4568985.zip", where 4568985 is your student ID.

Please try to make your submission file less than 20MB as you may have issues when uploading large files to MyAberdeen.

Any questions pertaining to any aspects of this assessment, please address them to the course coordinators Milan Markovic (milan.markovic@abdn.ac.uk) and Aiden Durrant (a.durrant.20@abdn.ac.uk).