COM6012 Assignment 2 - Deadline: 11:00 AM, Friday April 30, 2021

Please, carefully read the assignment brief before starting to complete the assignment

Assignment Brief

How and what to submit

A. Create a .zip file containing the following:

1) **AS2_report.pdf**: A report in PDF containing answers to ALL questions. The report should be concise. You may include appendices/references for additional information but marking will focus on the main body of the report.
2) **Code, script, and output files:** All files used to generate the answers for individual questions above, **except the data**. These files should be named properly starting with the question number: e.g., your python code as **Q2**_xxx**.py** (**one each question**), your script for HPC as **Q2**_HPC.sh, and your output files on HPC such as Q2_output.txt or Q2_figB.jpg. The results should be generated from the HPC, not your local machine.

B. Upload your .zip file to Blackboard (BB) before the deadline above. Name your .zip file as **USERNAME_AS2.zip**, where USERNAME is your username such as **abc18de**.

C. **NO DATA UPLOAD**: Please do not upload the data files used. We have a copy already. Instead, please use a **relative file path in your code (data files under folder 'Data')**, as in the lab sheets so that we can run your code smoothly.

D. **Code and output**. 1) Use **PySpark** as covered in the lecture and lab sessions to complete the tasks; 2) **Submit your PySpark job to HPC** with **qsub** to obtain the output.

Assessment Criteria (Scope: Session 6-9; Total marks: 20)

1. Being able to use pipelines, cross-validators and a different range of supervised learning methods for large datasets
2. Being able to analyse and put in place a suitable course of action to address a large scale data analytics challenge

Late submissions: We follow the Department's guidelines about late submissions, i.e., a deduction of 5% of the mark each working day the work is late after the deadline, but **NO late submission will be marked one week after the deadline** because we will release a solution by then. Please see this link.

Use of unfair means: "*Any form of unfair means is treated as a serious academic offence and action may be taken under the Discipline Regulations.*" (from the MSc Handbook). Please carefully read this link on what constitutes Unfair Means if not sure.

Please, only use interactive HPC when you work with small data to test that your algorithms are working fine. If you use rse-com6012 in interactive HPC, the performance for the whole group of students will be better if you only use up to four cores and up to 15G per core. When you want to produce your results for the assignment and/or want to request access to more cores and more memory, PLEASE USE BATCH HPC. This will be mandatory. We will monitor the time your jobs are taking to run and will automatically "qdel" the job if it is taken much more than expected. We want to promote good code practices (e.g. memory usage) so, please, once more, make sure that what you run on HPC has already been tested enough for a smaller dataset. It is OK to attempt to produce results several times in HPC, but please, be mindful that extensive running jobs will affect the access of other users to the pool of resources.

Question 1. Searching for exotic particles in high-energy physics using classic supervised learning algorithms [15 marks]

In this question, you will explore the use of supervised classification algorithms to identify Higgs bosons from particle collisions, like the ones produced in the Large Hadron Collider. In particular, you will use the HIGGS dataset.

About the data: "The data has been produced using Monte Carlo simulations. The first 21 features (columns 2-22) are kinematic properties measured by the particle detectors in the accelerator. The last seven features are functions of the first 21 features; these are high-level features derived by physicists to help discriminate between the two classes. There is an interest in using deep learning methods to obviate the need for physicists to manually develop such features. Benchmark results using Bayesian Decision Trees from a standard physics package and 5-layer neural networks are presented in the original paper. The last 500,000 examples are used as a test set."

You will apply Random Forests, Gradient boosting and (shallow) Neural networks over a subset of the dataset and over the full dataset. As performance measures use classification accuracy and area under the curve.

1. Use pipelines and cross-validation to find the best configuration of parameters for each model.
    a. For finding the best configuration of parameters, use 1% of the data chosen randomly from the whole set (2 marks).
    b. Use a sensible grid for the parameters (for example, three options for each parameter) for each predictive model (3 marks).
    c. Use the same splits of training and test data when comparing performances among the algorithms (1 mark).

    Please, use the batch mode to work on this. Although the dataset is not as large, the batch mode allows queueing jobs and for the cluster to better allocate resources.

2. Working with the larger dataset. Once you have found the best parameter configurations for each algorithm in the smaller subset of the data, use the full

dataset to compare the performance of the three algorithms in the cluster. **Remember to use the batch mode to work on this.**

   a. Use the best parameters found for each model in the smaller dataset of the previous step, for the models used in this step (2 marks)
   b. Once again, use the same splits of training and test data when comparing performances between the algorithms (1 mark)
   c. Provide training times when using 5 CORES and 10 CORES (2 marks). Based on our own solution, with the proper setting, you need a maximum of 10 mins for running the exercise when using 10 cores (with the rse-com6012 queue).

3. Report the three most relevant features according to each method in step 2 (1 mark). Note: you only need to do this for the ensemble methods, not for the neural network.
4. Discuss at least three observations (e.g., anything interesting), with one to three sentences for each observation (3 marks).

Do not try to upload the dataset to BB when returning your work. **It is 2.6Gb.**

COMMENTS: 1) To make sure you work with the same training and test data for all the different supervised methods, you can do randomSplit once and save the training and test sets in disk. **2)** An old, but very powerful engineering principle says: *divide and conquer*. If you are unable to analyse your datasets out of the box, you can always start with a smaller one, and build your way from it. **3)** This dataset was used in the paper "Searching for Exotic Particles in High-energy Physics with Deep Learning" by P. Baldi, P. Sadowski, and D. Whiteson, published in Nature Communications 5 (July 2, 2014). You can compare the results that you get against Table 1 of the paper. **4)** Use **wget** to download the data file directly in HPC.

Question 2. Senior Data Analyst at *Intelligent Insurances Co.* [15 marks]

In this Question we are going to look back at Assignment 1 from the module Machine Learning and Adaptive Intelligence (COM6509). You will revisit the example from that Assignment but now from the perspective of working with a larger dataset.

You are hired as a Senior Data Analyst at *Intelligent Insurances Co.* The company wants to develop a predictive model that uses vehicle characteristics to accurately predict insurance claim payments. Such a model will allow the company to assess the potential risk that a vehicle represents.
The company puts you in charge of coming up with a solution for this problem and provides you with a historic dataset of previous insurance claims. **The claimed amount can be zero or greater than zero and it is given in US dollars.** A more detailed description of the problem and the available historic dataset is here The website contains several files. You only need to work with the .csv file in **train_set.zip.** The uncompressed file is **2.66 Gb**. Back in COM6509 we only used a small subset of data from this file.

1. Preprocessing

    a. The dataset has several fields with missing data. Choose a method to deal with missing data (e.g. remove the rows with missing fields or use an imputation method) and justify your choice (1 mark).
    b. convert categorical values to a suitable representation (1 mark).
    c. the data is highly unbalanced: most of the records contain zero claims. When designing your predictive model, you need to account for this (2 marks).
2. Prediction using linear regression. You can see the problem as a regression problem where the variable to predict is continuous. Be careful about the preprocessing step above. The performance of the regression model will depend on the quality of your training data
    a. Use linear regression in PySpark as the predictive model. Partition your data into training and test (percentages according to your choice) and report the mean absolute error and the mean squared error (2 marks).
    b. Provide training times when using 5 CORES and 10 CORES. **Remember to use the batch mode to work on this** (1 mark). Based on our own solution, with the proper setting, you need a maximum of 5 mins for running the exercise when using 10 cores (with the rse-com6012 queue).
3. Prediction using a combination of two models. You can build a prediction model based on two separate models in tandem (one after the other). Once again, be careful about the preprocessing step above. For this step, use the same training and test data that you used in 2.a.
    a. The first model will be a binary classifier (of your choice) that will tell whether the claim was zero or different from zero. The performance of the classifier will depend on the quality of your training data (2 marks).
    b. For the second model, if the claim was different from zero, train a **Gamma regressor (a GLM)** to predict the value of the claim. Report the mean absolute error and the mean squared error (2 marks).
    c. Provide training times when using 5 CORES and 10 CORES. **Remember to use the batch mode to work on this** (1 mark). Based on our own solution, with the proper setting, you need a maximum of 15 mins for running the exercise when using 10 cores (with the rse-com6012 queue).
4. Discuss at least three observations (e.g., anything interesting), with one to three sentences for each observation (3 marks).

COMMENTS: 1) To make sure you work with the same training and test data for all the different supervised methods, you can do randomSplit once and save the training and test sets in disk. **2)** An old, but very powerful engineering principle says: *divide and conquer*. If you are unable to analyse your datasets out of the box, you can always start with a smaller one, and build your way from it. **3)** You may want to revisit your solution to the same problem when you used scikit-learn for Assignment 1 and perhaps apply here some lessons learned from solving the prediction problem there. **4)** You can download the data directly to your data folder in HPC using the following instructions:

1. Connect to ShARC and do qrshx
2. module load apps/python/conda

3. conda create -n mykaggle python=3.6
4. source activate mykaggle
5. pip install kaggle
6. Sign up for a Kaggle account at https://www.kaggle.com. Then go to the 'Account' tab of your user profile (https://www.kaggle.com/<username>/account) and select 'Create New API Token'. This will trigger the download of kaggle.json, a file containing your API credentials.
7. Copy kaggle.json to the cluster in your home folder in the location /home/<username>/.kaggle/kaggle.json
8. To protect your file from others, do
chmod 600 /home/<username>/.kaggle/kaggle.json
9. Type in your terminal
kaggle competitions download ClaimPredictionChallenge
All the files should download in seconds.

If you need more information about the Kaggle API, you can find it in this link.