Escuela Colombiana de Ingenieria

MSc in Data Science

Title: Application of Machine Learning Methods to Real-World Data

Module Name:	Machine Learning
Module Code:	MLEA_M
Level:	MSc.
Programme:	MSc Data Science
Type of assessment:	Coursework
Weighting:	100%
Max. mark available:	50

Lecturer:	Dr Ivan Olier-Caparroso
Contact details:	ivan.olier@escuelaing.edu.co

Resource requirements: Desktop/Laptop computer, Module Notes, Python, Microsoft Word, Library Resources, Internet.

Important dates:

Hand-out date:	04 Nov 2022

InClass Kaggle competition

Closing date:	03 Dec 2022, 23:59 (UTC)

Coursework submission

Hand-in date:	05 Dec 2022, 23:59 (UTC)
Hand-in method:	Microsoft Teams

Feedback date:	09 Dec 2022
Feedback method:	On Microsoft Teams

Introduction

This coursework provides experience in using the methods developed theoretically in class. In particular, you will be provided with a real-world problem and be asked to provide a solution using machine learning.

Coursework format

This coursework requires you to work individually. You are required to submit a brief report that summarises the work done and the predictions of the final model you develop that you consider the best possible one.

This is not prescriptive coursework with a clear path to the solution. Instead, it requires you to conceive, code and test several approaches before you reach a final solution. Also, training ML models require a certain amount of computing time that may further slow your progress. Therefore, it is highly unwise to leave the work to the last minute. It is also expected that a significant amount of the work will be carried out during the subsequent IT lab sessions.

As part of the coursework assessment, a leader board will be created. This requires you to submit your predictions in a standard file format. See details in the "**What you need to submit**" section.

Details of the real-world problem

Email spam, also referred to as junk email, spam mail, or simply spam, is unsolicited messages sent in bulk by email (spamming). The name comes from a Monty Python sketch in which the name of the canned pork product Spam is ubiquitous, unavoidable, and repetitive. Email spam has steadily grown since the early 1990s, and by 2014 was estimated to account for around 90% of total email traffic.

The aim of this coursework is to develop a solution based on Machine Learning to predict whether an email is likely to be spam or not based on already extracted features. The coursework's database is extracted from the UCI Machine Learning repository. However, notice that I have made significant changes to the database to make it suitable for this coursework. Therefore, you will certainly be at risk of plagiarism if your submitted coursework mainly uses code already published (see more details under the Academic Misconduct section of this assignment). However, you can surf the Internet to find ideas on how to solve the coursework. If you have any doubt, please discuss with me the situation as soon as possible but always before submission.

Database description

The coursework's database is already available on MS Teams (it is also available on the InClass Kaggle Competition). It is the official database to be used in this coursework, and no other variant that could be available elsewhere will be allowed to be used.

Description of the files and folders

- spam_dataset.csv - Tabular dataset for the training set
- spam_dataset_test.csv - Tabular dataset for the test set
- predictions_specimen.csv - A sample submission file in the correct format

Data fields

Most of the attributes in "spam_dataset.csv" indicate whether a particular word or character was frequently occurring in the email. The run-length attributes (55-57) measure the length of sequences of consecutive capital letters. Here are the definitions of the attributes:

email_id continuous integer [0,…]. It it the row identifier, which must not be included in the models.

48 continuous real [0,100] attributes of type word_freq_WORD = percentage of words in the e-mail that match WORD, i.e. 100 * (number of times the WORD appears in the e-mail) / total number of words in e-mail. A "word" in this case is any string of alphanumeric characters bounded by non-alphanumeric characters or end-of-string.

6 continuous real [0,100] attributes of type char_freq_CHAR = percentage of characters in the e-mail that match CHAR, i.e. 100 * (number of CHAR occurences) / total characters in email.

1 continuous real [1,...] attribute of type capital_run_length_average = average length of uninterrupted sequences of capital letters.

1 continuous integer [1,...] attribute of type capital_run_length_longest = length of longest uninterrupted sequence of capital letters.

1 continuous integer [1,...] attribute of type capital_run_length_total = sum of length of uninterrupted sequences of capital letters = total number of capital letters in the e-mail.

1 nominal {0,1} class attribute of type spam = denotes whether the e-mail was considered spam (1) or not (0), i.e. unsolicited commercial e-mail.

spam column {1,0} that indicates whether the email was spam or not. This is the target variable.

The "spam_dataset_test.csv" file contains all the above fields but "spam", which is unknown for the test set.

Outcome

The main coursework outcome is to predict the risk of an spam email.

What you need to submit

On Teams, you will have two separate assignments: 1) to submit your model predictions, and 2) to submit your final report:

Submission of model predictions

You must run your model(s) on the supplied test data and submit the predicted outcome in the form of probabilities. You must use the file format as in the *predictions_example.csv* file, which is available on Kaggle. Your file name must be "**predictions_XXXX.csv**", where XXXX is the name of your team. Important: Note that you won't be awarded any marks associated with the *Model predictions* assessment component if you fail to submit this file or to submit it using an incorrect format.

Your submitted final predictions will be used to generate a leader board based on model performance on the test data as measured using the AUC (area under the ROC curve). In order to avoid overfitting the test set, this is randomly split into two subsets of similar size. The AUC is calculated on both data splits. The student with the highest average AUC on the test splits will win the competition.

You will receive marks based on your position on the leaderboard, which will be estimated using the most recent submission after the submission system is shut down. Although you are allowed to submit as many prediction files as you want through MS Teams, it is your responsibility to verify that the final submission is the one you want to use for the official leaderboard.

There is a *Community Kaggle Competition* associated with this assessment that you could use to test your predictions and to see how they compare against your classmates. I strongly recommend the use of this facility. See more information in the *Community Kaggle Competition* section.

Submission of the final report

You must produce a report that summarises your main results. Although the main content of your report should be the presentation and discussion of your results, you should also describe how you addressed the task, alternative solutions, possible reasons for success/failure. Also, a brief reflection on your work should be included. You must list your code as an appendix.

I suggest structuring your report as follows:

1. Description of approaches used to solve the problem (indicating final and alternative approaches, methods, portions of the data used, etc)
2. Results (of the final and alternative approaches, including ROCs, performance tables, etc.)
3. Discussion of the results (explaining reasons for success/failure of the considered approaches)
4. Reflection (insights for future improvements)

   Appendix 1 – Code listings

Excluding appendices, it is expected the report length to be between 2500 and 4000 words (5 – 8 pages). You must submit one file only in PDF/DOC/DOCX format only (preferably PDF). If you submit more than one file, only the first one will be marked. You can use any word processor, provided that you manage to export your report to any of the acceptable file formats.

Community Kaggle Competition (https://www.kaggle.com/c/MLEAM222)

This assignment has an associated *Community Kaggle* competition. You can make use of this facility to test as many prediction files as you want. The competition is already open and will close several days before your coursework's submission deadline. Invitations to join the competition will be shared via Teams. Important: Note that submitting predictions to Kaggle does not constitute a formal submission to any of the assessment components. You must still submit your prediction file through MS Teams. The *Community Kaggle Competition* is available to help you to decide which prediction file you submit based on how you rank against your classmates. Also, the only valid leaderboard will be the one generated from the prediction files submitted to MS Teams. You might expect minor differences between both leader boards, the one generated by Kaggle and the one generated from your submissions. The Kaggle's competition will also split the test set so two leaderboards are produced, public and private. The public leaderboard will be visible to you all the time via Kaggle. However, the private leaderboard will be only available during the IT lab sessions.

Assessment Criteria

The coursework is 100% of the assessment for this module. It will be marked out of 50. The breakdown of the marks available is as follows:

- Report – up to 40
- Model predictions – up to 10

Please refer to the Appendix for the marking rubric to be used to assess your coursework.

Any coursework submitted late without my approval will receive 0 marks.

Academic Misconduct

The University defines Academic Misconduct as ‘any case of deliberate, premeditated cheating, collusion, plagiarism or falsification of information, in an attempt to deceive and gain an unfair advantage in assessment’. This includes attempting to gain marks as part of a team without contributing. The Faculty takes Academic Misconduct very seriously and any suspected cases will be investigated through the University’s standard policy. If you are found guilty, you may be expelled from the University with no award.

It is your responsibility to ensure that you understand what constitutes Academic Misconduct and to ensure that you do not break the rules. If you are unclear about what is required, please ask.

Appendix – Marking rubric

Rubric used to assess final model predictions (marks)

Criteria	10 out of 10	9 out of 10	8 out of 10	6 out of 10	4 out of 10	2 out of 10	0 out of 10
Model performance on the test set as measured using the area under the ROC curve (AUC)	Submitted prediction file correctly formatted. Leader board position 1.	Submitted prediction file correctly formatted. Leader board position 2.	Submitted prediction file correctly formatted. Leader board position 3 or 4.	Submitted prediction file correctly formatted. Leader board position 5 to 8.	Submitted prediction file correctly formatted. Leader board position 9 to 14.	Submitted prediction file correctly formatted. Leader board position 15 or below.	No prediction file was submitted, or it was not usable (i.e. wrongly formatted)

Rubric used to assess the report (marks)

Criteria	40	37	35	30	27	23	15	9	0
Description of the approaches used to solve the problem (up to 6 marks)	Extraordinary with several paths to a solution which are innovative and beyond the current state-of-the-art. **[6 marks]**	Excellent with several paths to a solution which are sophisticated and convincing. **[5 marks]**		Good with congruent and consistent paths to a solution. **[4 marks]**		Descriptive with unsophisticated paths to a solution. **[3 marks]**	Inadequate and/or contradictory. Paths to a solution are vaguely described. **[2 marks]**	Erroneous, insufficient and/or inappropriate description. **[1 mark]**	Missing or unrelated to the problem. **[0 marks]**
Results (up to 18 marks)	Description of the results is extraordinary. Analysis is performed using a vast range of models. Models are validated and compared in many ways. **[18 marks]**	Description of the results is outstanding. Analysis is performed using a vast range of models. Models are validated and compared in many ways. **[17 marks]**	Description of the results is Excellent. Analysis is performed using a handful of models. Models are validated and compared in several ways. **[16 marks]**	Description of the results is fluent. Analysis is performed using a handful of models. Models are validated and compared in several ways. **[14 marks]**	Description of the results is good. Analysis is performed using a handful of models. Models are validated and compared in several ways. **[12 marks]**	Description of the results is adequate. Analysis is performed using a few models. Limited model validation and comparison. **[10 marks]**	Inadequate details of the results are provided and supported. Analysis is limited to one model. Models are not properly validated and compared. **[6 marks]**	Erroneous details on the results are provided and supported. Analysis is limited to one model. Models are not properly validated and compared. **[3 marks]**	Results are missing. Very limited evidence that the analysis was performed. **[0 marks]**

Discussion of the results (up to 12 marks)	Discussion of the results is exceptional and clearly distinctive. **[12 marks]**	Discussion of the results is outstanding and insightful. **[11 marks]**	Discussion of the results is excellent critical. **[10 marks]**	Discussion of the results is credible and precise. **[9 marks]**	Discussion of the results is accurate and coherent. **[8 marks]**	Discussion of the results is adequate. **[7 marks]**	Discussion of the results is imprecise, limited and/or inadequate. **[4 marks]**	Discussion of the results is ambiguous, incoherent, irrelevant and/or erroneous. **[2 marks]**	Discussion of the results is missing, or unrelated to the results or the problem. **[0 marks]**
Reflection (up to 2 marks)	Exceptional evaluation and distinctive reflection on lessons learnt. **[2 marks]**			Good evaluation and congruent reflection on lessons learnt. **[1 mark]**					Reflection is missing or unrelated to the problem. **[0 marks]**
Code listings (up to 2 marks)	Code is attached and seems to provide a clear path to a solution to the problem. **[2 marks]**								Code is missing, or so limited as to provide no clear path to a solution to the problem. **[0 marks]**