CM20315 – Machine Learning

Coursework 1: Regression Performance

Set: 22/10/2024 (week 4) **Due: 15/11/2024 (week 7), 8pm**
Percentage of overall unit mark: **15%**
Submission Location: Moodle **Submission Components: Poster** **Submission Format: 1 x PDF file and 1 x Jupyter Notebook.**
Anonymous Marking: Y

1 Overview

This coursework assesses two of the three module Learning Outcomes:

2. Demonstrate understanding of a wide range of machine learning techniques, their strengths and their limitations.
3. Write code in a relevant programming language and employ software libraries to solve problems in machine learning.

For this coursework you will be selecting, implementing, and evaluating regression models for a scenario of your choice.

1.1 The Task

You must complete the following tasks, documenting your process and findings in the form of a poster (as defined in section 2):

- Select a scenario and research question from the options given in section 6.
- Train at least two different machine learning models to answer your selected research question.
- Evaluate your machine learning models using suitable criteria.
- Conclude by providing a recommended machine learning approach to the research question.

2 Deliverable

There are two deliverables to this coursework:

Deliverable	Format	Content
A3 Poster	.pdf file	Overview of your experiments, training process, evaluation and final recommendation.
Code	.ipynb	The code necessary to run your experiments, train the models and produce necessary graphics.

The poster must contain the following subsections:

Section	Expected Content
Introduction	Overview of the problem and poster: • What is this poster showing? • What was the primary research question?
Methodology	Details of the methodology employed: • What is your modelling approach? • What data is being used to train, test and validate each model? • What hyperparameters have you selected and why? • What is your approach to evaluation?
Results	Details of the results from training and evaluation: • How do the models compare? • Are there insights about model behaviour you can draw on?
Conclusions	Overall conclusions from the experiments and your recommendation: • What were the key findings and results? • What were the limitations of your study? • What could you do differently to improve the study?

Table 1. Poster contents.

2.1 Mark Scheme

Your poster will be assessed using the contributions below, each section will be marked according to the marking grid provided at the end of this document. Whilst the code itself does not form part of the marking criteria it is required for evidence that the results are indeed true.

Criteria	Percentage of overall deliverable mark
Introduction	10%
Methodology	30%
Results	20%
Conclusions	30%
Presentation	10%

Table 2. Mark Allocation.

3 Feedback

Formative feedback will be available in the labs.

You will receive **summative feedback** on your work within 3 semester weeks of the submission deadline. The feedback will discuss your performance based on the criteria for marking, including what you did well and how specific components/sections could have been improved.

4 Academic Integrity

Your work will be checked to ensure that you have not plagiarised. For more information about the plagiarism policy at the University see: https://library.bath.ac.uk/referencing/plagiarism

Remember that published work that you refer to in your poster should be clearly referenced in your text and listed in a bibliography section given at the bottom of your poster. For more information see, https://library.bath.ac.uk/referencing/new-to-referencing

5 FAQ

5.1 Is the quality of my models marked?

In short, no. We are more interested in whether you can adequately explain and evaluate your models' performances using the metrics and evaluation methods you choose. There are no direct marks available for the quality of the model.

5.2 Why do I only have one side of A3?

When faced with these challenges in the real world it is important that you can provide succinct but

detailed reports to clients. There is no room for flamboyant language or unnecessary detail, I suggest you are creative in how you use the size limit with figures and tables where necessary, the less figures you need to use, the more space you have but sometimes a picture can paint a thousand words.

5.3 Do we have to write the algorithms ourselves?

No! The point of this exercise is to allow you the freedom to select and implement necessary methods from Python libraries, whilst you may wish to write the algorithms yourself this would have not fluency on your mark. Consider the task, the short time frame and remember this is primarily about your knowledge of evaluating and interpreting machine learning models using appropriate techniques, not how well you can program in Python.

5.4 How many methods do we need to evaluate?

To make a comparison you'll need at least 2 and there is no upper limit. However, think carefully again about the space given to report your results and the time you need to perform the evaluation.

5.5 Can we use methods that haven't been taught?

If you wish to use methods or validation metrics which you have found independently or used last semester you may use these if you make the adequate justifications in the poster. This will not attract more marks.

5.6 Does my poster have to be portrait or landscape?

You are free to use either portrait or landscape as long as it is A3.

5.7 Do you have any advice for creating a poster?

Posters should be eye-catching, straight to the point and graphical. For formal advice, the university has a blog on good poster tips: https://blogs.bath.ac.uk/technicians/2019/05/15/poster-writing-hints-and-tips/. Here is another useful resource: https://www.animateyour.science/post/how-to-write-engaging-headings-to-make-your-scientific-poster-pop.

5.8 How do I create an A3 poster?

The easiest way to create an A3 poster (and export to PDF) is by using Microsoft PowerPoint:

1. Open an empty PowerPoint with the starting slide.
2. Go to the 'Design' tab.
3. Click on 'Slide Size' (usually on the right-hand side of the toolbar).
4. Click 'Custom Slide Size…'.
5. Select A3 Paper.

Once you've finished your poster you can use the export function in the 'File' menu to save as PDF.

6 Scenarios

For this coursework you may choose from the following three scenarios:

Scenario	**Rainfall Forecasting**	**Predicting Forest Fire Extent**	**House Price Prediction**
Research Question	Can we predict monthly rainfall given forecasted minimum and maximum temperatures?	Can we predict the extent of a forest fire given meteorological conditions?	Can we predict house prices based on given features?
Description	Given the increasing threat to society caused by extreme weather events a new investigation has been launched into improving monthly weather forecasts. Forecasting the minimum and maximum temperatures of a given month is already possible thanks to mathematical models; however, rainfall forecasts are severely lacking and many hope machine learning could be the solution. You have been brought in to carry out a series of experiments using machine learning to forecast rainfall given the minimum and maximum temperature forecasts. Historic data for three weather stations in Wales have been provided for you.	Forest fires are becoming increasingly common in southern Europe and the extent of these forest fires can have drastic consequences on local societies and economies. Being able to predict the potential extent of forest fires can be crucial to enabling emergency services to take appropriate measures. You have been given a dataset of historic forest fires for a park in Portugal (Montesinho park) and have been asked to develop a regression model for predicting the extent of forest fires in the park given the range of features provided to you.	Estimating the price of a house prior to putting it on the market can be a critical task for estate agents to enable expectation management of their customers. You have been provided a dataset with three key features to house prices in central London along with the final selling price of the house on a per metre squared basis. You have been asked to develop a model for estimating the sale price (on a per metre squared basis) of a house given the features provided in the dataset.

Dataset	The dataset provided is in the form of three Numpy files (one for each weather station): • Cardiff.npy • Aberporth.npy • Valley.npy Each dataset contains 6 columns as follows: • Year: Year of reading. • Month: Month of reading. • Minimum Temperature: Minimum temperature forecast for the month in Celsius. • Maximum Temperature: Maximum temperature forecast for the month in Celsius. • No. Frost Days: The number of days forecast for given the month. • Monthly Rainfall: The amount of rainfall which fell in the month in mm. Each row represents one month. To load the dataset using numpy you can use the following command: dataset = np.load("<FILENAME>")	The dataset provided to you is in a single CSV file: • montesinho.csv The dataset contains 6 columns: • DMC Index: A numerical rating based on the moisture content of compact organic matter. • Temperature: Temperature of the park in Celsius. • Relative Humidity: Relative humidity as a %. • Wind: Wind speed in km/h. • Rainfall: Total rainfall in mm/m^2 • Area: The total area of the forest which burned (in hectares). Each row represents one reading. To load the dataset using numpy you can use the following command: dataset = np.loadtxt("montesinho.csv", delimiter=',')	This dataset is provided to you in a single CSV file: • realestate.csv The dataset contains 4 columns: • House Age: Years since the house was built. • Distance to Tube: The distance to the closest tube station in metres. • Nearby Stores: The number of stores within a 2km radius. • Price per m^2: the price the house sold for per m^2. Each row represents one previous house sale. To load the dataset using numpy you can use the following command: dataset = np.loadtxt("realestate.csv", delimiter=',')

Marking Rubric

Section	%	Mark				
		Fail	3:3	2:2	2:1	1:1
Introduction	10%	No introduction provided.	Limited introduction.	Moderate introduction addressing the key elements required.	Good, sharp introduction with suitable detail of all key elements.	Excellent introduction detailing all required components in a succinct manor.
Methodology	30%	No or limited description of the methodology.	Methodology provided with incomplete or missing information.	Primary components of the methodology presented to a good standard with some missing components.	All components of the methodology are addressed and described.	All components of the methodology are addressed, described and adequately justified where necessary.
Results	20%	No or limited results presented.	High level results presented with limited interpretation.	High-level results presented with some superficial interpretation.	Good selection of results shown in a clear and meaningful way with a fair level of interpretation.	Clear, succinct evaluation of the results with distinctive interpretation given to all aspects of differentiation.
Conclusions	30%	Limited or no conclusions drawn from results.	Limited, high-level conclusions drawn with limited links to the results section and some minor reflection.	Good range of conclusions drawn with some links to the results and an adequate study evaluation.	Good range of conclusions with clear supporting evidence from the results and an adequate evaluation of the study considering some future work.	Excellent range of conclusions drawn with distinct supporting evidence. The study has been evaluated and reflections are clear with relevant future work identified.
Presentation	10%	Poster is presented with limited structure, has substantial errors or uses text and images which are unreadable.	The poster has an evident structure but with unconsidered use of images, tables and text.	The poster has a good structure with appropriate use of images, tables and text.	The poster has a good structure with a good use of images, tables and text where appropriate. The poster has an evident but limited flow.	The poster has an exceptional structure with excellent use of images, text and tables. The flow of the poster is uninterrupted.