COMP3411/9814 24T3
Artificial Intelligence

Assignment 1 - Artificial neural networks

Due: Week 5, Friday, 11 October 2024, 5pm.

1 Problem overview

In this assignment, you will use artificial neural networks for drought modelling in the Murray-Darling Basin. You will conduct two tasks:

(a) A classification task to predict whether there is 'a drought' or 'no drought' based on the climate conditions.

(b) A regression task to predict the intensity of a drought based on the climate conditions.

The Murray-Darling Basin is a very important agricultural region in Australia that produces a large portion of the nation's food and fibre (see Fig. 1). It is a big river system that supports many different types of crops and livestock, making it essential for Australia's food production. However, the basin is very prone to droughts, which can greatly affect water availability and agricultural productivity, threatening the stability of the food system. Because of this, there is a lot of interest in monitoring drought conditions in the Murray-Darling Basin to manage water resources effectively.

Figure 1: Location of the Murray-Darling Basin and illustration of grid cells within the basin.

2 Dataset

The climate data used in this assignment is derived from the ERA5 climate dataset. ERA5 data covers the globe and combines surface observations, satellite measurements, and weather balloons, using advanced numerical weather prediction models and data assimilation methods. The data you will use has been extracted from the ERA5 global dataset for the Murray-Darling Basin in Australia. The data has been processed to provide monthly values of several climate variables associated with droughts and covering the years from 1979 to 2020. The original ERA5 dataset is publicly available and consists of more than 200 terabytes. However, the datafile provided for this assignment in `Climate_SPI.csv` has been specifically derived and cannot

be downloaded from anywhere else.

Table 1: List of variables included in data file Climate_SPI.csv.

Variable	Description
mn2t	Minimum temperature at 2 meters (°K)
msl	Mean sea level pressure in (Pa)
mx2t	Maximum temperature at 2 meters (°K)
q	specific humidity (kg kg-1)
t	Average temperature at Pressure level 850 hPa in (°K)
t2	Average temperature at 2 m in (°K)
tcc	Total cloud cover (0-1)
u	U wind component at pressure level 850 hPa (ms-1)
u10	U wind component at 10 m (ms-1)
v	V wind component at Pressure level 850 hPa (ms-1)
v10	V wind component at 10 m (ms-1)
z	Geopotential (m2s-2)
month	1 to 12
year	1979 to 2020
grid_ID	The ID of the grid cell
SPI	Standardised Precipitation Index (unitless)

2.1 Climate predictors and additional attributes

The data file `Climate_SPI.csv` contains climate data for 30 grid cells. Each grid cell is represented as a single square within the boundaries of the basin, as shown in Fig. 1. In the data file, each row, or sample, corresponds to a grid cell identified by grid_ID, and contains:

- Time information: year, month.
- A set of climate variables represented by acronyms: mn2t, msl, mx2t, q, t, t2, tcc, u, u10, v, v10, and z. The description and unit of measurement for each of these variables are provided in Table 1. A time series plot of each of these variables for a single grid is shown in Fig. 2 to help you visualise these variables.
- A drought index: SPI.

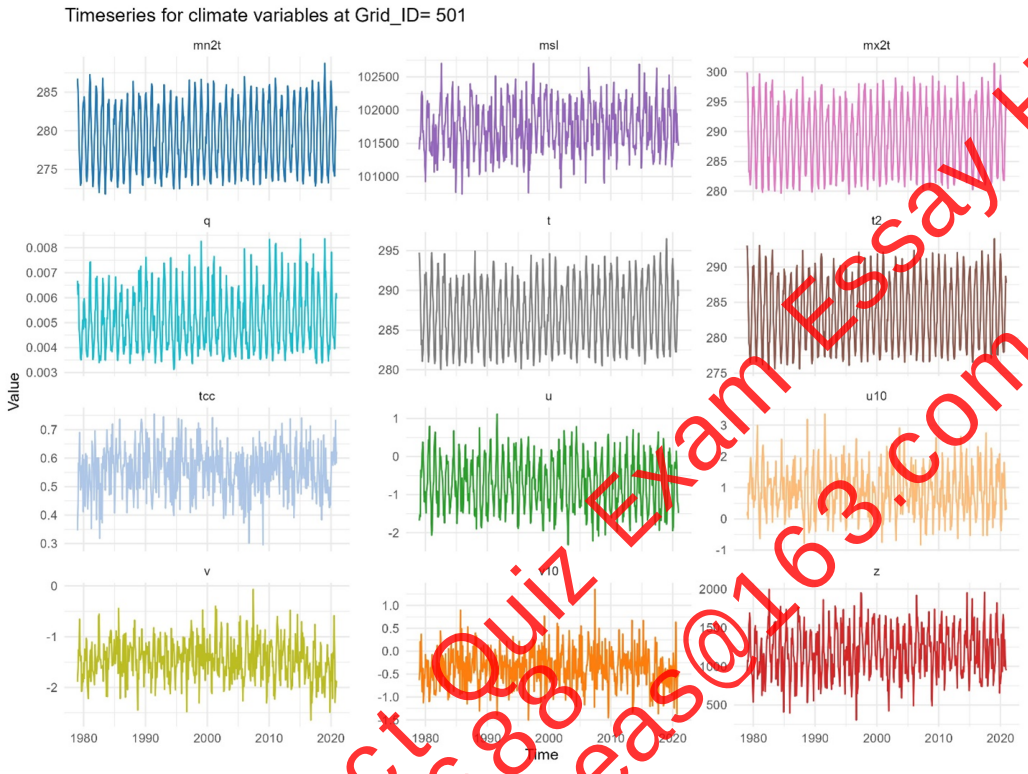

Figure 2: Time series of the climate variables for a single grid cell.

2.2 Target variable

The drought index SPI (Standardised Precipitation Index) is used here as a proxy to characterise the intensity of drought caused by precipitation deficiency. Very low SPI values suggest intense drought, while very high values indicate very wet conditions. In the regression task, SPI is your target variable, and you need to predict it based on the climate variables. In the classification task, you will calculate a binary target variable 'Drought' from SPI. The 'Drought' variable will be 1 to indicate the occurrence of drought and 0 to indicate no drought. You will apply a threshold of -1 to SPI, where values below or equal to this threshold indicate drought conditions (i.e. Drought = 1), otherwise, there is no drought (i.e. Drought = 0). A time series plot of SPI for a single grid cell is shown in Fig. 3. Periods where SPI falls below or is equal to -1 indicate periods of drought.

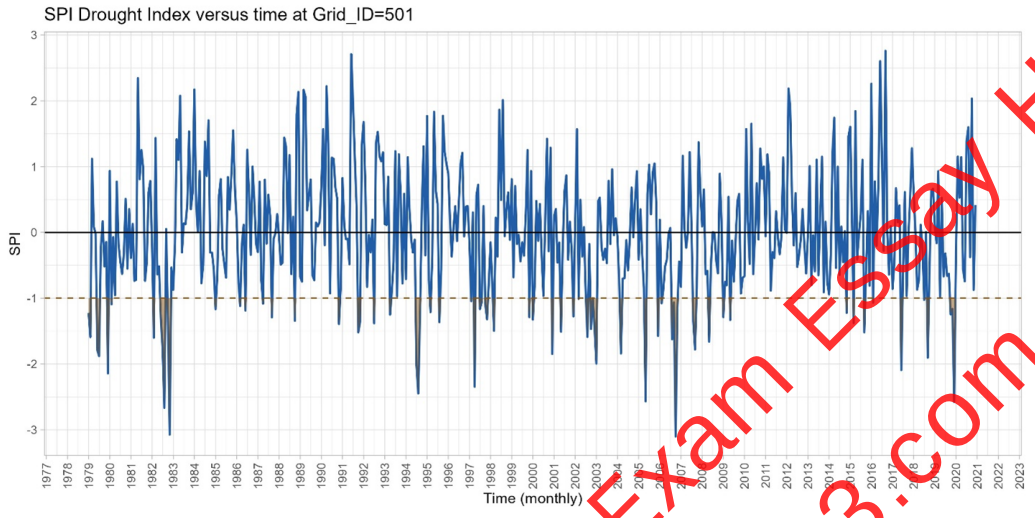

Figure 3: Time series of the drought index SPI for a single grid cell. Periods where SPI falls below or is equal to -1 indicate periods of drought.

2.3 Data cleansing

Next is the description of the classification and regression tasks. You can treat each task separately. Do not use 'year' and 'Grid_ID' for either task. Additionally, you cannot use 'Drought' and 'SPI' as predictors for either task. For both tasks, you will need to create your artificial neural network using Keras and TensorFlow packages. Before you start solving the tasks, please carefully read the entire assignment description, including how to submit your assignment.

3 Classification task

In this classification task, the goal is for the neural network to predict whether there is a drought or no drought based on climate data predictors. A complete script should include all the steps below, along with any additional necessary steps:

(a) Calculate 'Drought' from 'SPI' by setting a threshold of -1 for the 'SPI' values.

(b) Split your data into training, validation and test sets.

(c) Pre-processing: Apply any necessary transformation to the training set, then apply the same transformation to the validation and test sets. Keep record of all applied transformations.

(d) Build your model by defining its architecture, and hyperparameters. This includes: loss function, optimiser, batch size, learning rate, and number of epochs. It is recommended that your network has fewer parameters than the number of samples divided by 10.

(e) Train the neural network and monitor the evolution of the loss values.

3.1 Evaluate model performance on validation set

(f) Create a plot showing the accuracy (y-axis) versus the number of epochs (x-axis) for both the training and validation sets.

3.2 Optimise your neural network

(g) Repeat steps d), e), and f) to find the best neural network configuration.

(h) Build and train your model on different subsets of climate predictors. Use the subset of climate predictors that achieve the best outcome based on the validation set.

3.3 Evaluate your classification model on test set

(i) Apply the same transformations to the test set as you did to the training set.

(j) Use your model to predict the class on the test set.

(k) Evaluate the performance of your model using sklearn and/or creating your own functions:

- Compute and plot a confusion matrix. Note that your positive class is 1, i.e. 'Drought'.
- Calculate and print "Balanced Accuracy" and "Precision".

3.4 Additional notes (apply to both tasks)

- You need to set the random seed to ensure that your results are reproducible.
- You must save your model.
- In step f), you only need to show the plot(s) for the best neural network configuration and the chosen predictors.
- If you choose to use 'month' as a predictor, encode it using cyclic encoding to ensure that your neural network understands the relationship between December and January. To achieve this:
  - Normalise the month to the range $[0, 2\pi]$ using: month_normalised $= 2\pi \times$ (month - 1)/12.
  - Replace 'month' with two new predictors: 'cos(month_normalised)' and 'sin(month_normalised)'.

4 Regression task

In this regression task, the goal is for the neural network to predict the intensity of the drought, represented by 'SPI' based on climate data predictors. A complete script should include all the steps below, along with any additional necessary steps.

(a) Split your data into training, validation and test sets.

(b) Pre-processing: Apply any necessary transformation to the training set, then apply the same transformation to the validation and test sets. Keep record of all applied transformations.

(c) Build your model by defining its architecture, and hyperparameters. This includes: loss function, optimiser, batch size, learning rate, and number of epochs. it is recommended that your network has fewer parameters than the number of samples divided by 10.

(d) Train the neural network and monitor the evolution of the loss values.

4.1 Evaluate model performance on validation set

(e) Assess the model's performance over the epochs by creating a plot showing the loss value (y-axis) versus the number of epochs (x-axis) for both the training and validation sets.

4.2 Optimise your neural network

(f) Repeat steps c), d), and e) to find the best neural network configuration.

(g) Build and train your model on different subsets of climate predictors. You may start with the same set of predictors that achieved the best outcome in the regression task.

4.3 Evaluate your regression model on test set

(h) Apply the same transformations to the test set as you did to the training set.

(i) Use your model to predict SPI on the test set.

(j) Evaluate the performance of your model using sklearn and/or scipy metrics.

- Create a scatter plot showing predicted SPI (y-axis) versus true SPI (x-axis).
- Calculate and print "Mean Absolute Error (MAE)" and the "Pearson Correlation Coefficient" between the true and predicted SPI.

5 Evaluating your work on a new dataset

In addition to the scripts you wrote for the classification and regression tasks, you need to prepare a script to evaluate your model on new data. The new data will be formatted exactly like `Climate_SPI.csv`, so you can assume `Climate_SPI.csv` is the new data for now. Your script should include all the steps below, along with any additional necessary steps.

5.1 Evaluating your classification model on a new dataset

(a) Load `Climate_SPI.csv`.

(b) Calculate 'Drought' from 'SPI' by setting a threshold of -1 for the 'SPI' values, where values below or equal to this threshold indicate drought conditions (i.e. Drought = 1); otherwise, there is no drought (i.e. Drought = 0).

(c) Apply the same transformations to the new data as you did to the training set in the classification task.

(d) Load your classification model.

(e) Use your model to predict the class 'Drought' on the new data.

(f) Compute and plot a confusion matrix.

(g) Calculate and print "Balanced Accuracy" and "Precision".

(h) Print the number of samples and your model's predictors set.

You must demonstrate a complete understanding of the code and analysis during the discussion.

5.2 Evaluating your regression model on a new dataset

(a) Load `Climate_SPI.csv`.

(b) Apply the same transformations to the new data as you did to the training set in the regression task.

(c) Load your regression model.

(d) Use your model to predict 'SPI' on the new data.

(e) Create a scatter plot to show predicted SPI (y-axis) versus true SPI (x-axis).

(f) Calculate and print "Mean Absolute Error" and the "Pearson Correlation Coefficient" between the true and predicted SPI.

(g) Print the number of samples and your model's predictors set.

You must demonstrate a complete understanding of the code and analysis during the discussion.

6 Testing and discussing your code

As part of the assignment evaluation, your code will be tested by tutors along with you in a discussion session carried out in the tutorial session. The assignment has a total of 25 marks. The discussion is mandatory and, therefore, we will not mark any assignment not discussed with tutors.

You are expected to propose and build neural models for classification and regression tasks. You will receive marks for each section as shown in Table 2. For marking your results, you should be prepared to simulate your neural model with a generalisation set we have saved apart for that purpose.

You will receive 1 mark for code readability, and your tutor will also give you a maximum of 8 marks for each task depending on the level of code understanding as follows: **8. Outstanding, 6. Great, 4. Fair, 2. Low, 0. Deficient or No answer**.

7 Submitting your assignment

The assignment must be done individually. You need to submit your solution on Moodle. Your submission must include:

- A single Jupyter notebook (.ipynb).
- The trained models for both the classification task and the regression task.

The first line of your Jupyter notebook should display your full name and your zID as a comment. The notebook should contain all the necessary code for reading files, data preprocessing, network architecture, and result evaluations. Additionally, your file should include short text descriptions to help markers better understand your code. Please be mindful that providing clean and easy-to-read code is a part of your assignment.

You can submit as many times as you like before the deadline – later submissions overwrite earlier ones. After submitting your file a good practice is to take a screenshot of it for future reference.

Table 2: Marking scheme for the assignment.

Criteria	Marks
Classification Task	
Plot of the accuracy (y-axis) versus the number of epochs (x-axis) for both the training and validation sets.	1 mark
Performance metrics "Balanced Accuracy" and "Precision" calculated on the test set.	1 marks
Confusion matrix on the unseen data.	1 mark
"Balanced Accuracy" and "Precision" on the unseen data.	1 marks
Demonstrate complete understanding of the code and analysis during discussion.	8 marks
Regression Task	
A plot showing the loss value (y-axis) versus the number of epochs (x-axis) for both the training and validation sets.	1 mark
Create a scatter plot showing predicted SPI (y-axis) versus true SPI (x-axis) based on the test set.	1 mark
"Mean Absolute Error (MAE)" and the "Pearson Correlation Coefficient" between the true and predicted SPI.	1 marks
A scatter plot showing predicted SPI (y-axis) versus true SPI (x-axis) based on the unseen dataset.	1 mark
Demonstrate complete understanding of the code and analysis during discussion.	8 marks
Code understanding and discussion	
Overall code readability including tidy and well-commented script	1 mark
Total marks	25 marks

Late submission penalty: UNSW has a standard late submission penalty of 5% per day from your mark, capped at five days from the assessment deadline, after that students cannot submit the assignment.

8 Deadline and questions

Deadline: Week 5, Friday 11 October 2024, 5pm. Please use the forum on Moodle to ask questions related to the assignment. We will prioritise

questions asked in the forum. However, you should not share your code to avoid making it public and possible plagiarism. If that's the case, use the course email `cs3411@cse.unsw.edu.au` as alternative.

Although we try to answer questions as quickly as possible, we might take up to 1 or 2 business days to reply, therefore, last-moment questions might not be answered timely.

For any questions regarding the discussion sessions, please contact directly your tutor. You can have access to your tutor email address through Table 3.

Table 3: COMP3411/9814 24T2 Tutorials

Number	Class ID	Time	Tutor	Email
1	4106	Wed 16:00 - 18:00	Ramya Kumar	ramya.kumar1@student.unsw.edu.au
2	4107	Wed 18:00 - 20:00	Janhavi Jain	j.jain@student.unsw.edu.au
3	4101	Thu 11:00 - 13:00	Maryam Hashemi	m.hashemi@unsw.edu.au
4	4102	Thu 11:00 - 13:00	Maher Mesto	m.mesto@unsw.edu.au
5	4103	Thu 13:00 - 15:00	Maryam Hashemi	m.hashemi@unsw.edu.au
6	6138	Thu 17:00 - 19:00	Ramya Kumar	ramya.kumar1@student.unsw.edu.au
7	6139	Thu 19:00 - 21:00	Maher Mesto	m.mesto@unsw.edu.au
8	4097	Fri 11:00 - 13:00	Zhijin Meng	zhijin.meng@student.unsw.edu.au
9	6132	Fri 13:00 - 15:00	Zhijin Meng	zhijin.meng@student.unsw.edu.au
10	4099	Fri 15:00 - 17:00	Janhavi Jain	j.jain@student.unsw.edu.au
11	6134	Fri 17:00 - 19:00	Stefano Mezza	s.mezza@unsw.edu.au
12	13044	Thu 18:00 - 20:00	John Chen	xin.chen9@student.unsw.edu.au
13	12705	Wed 16:00 - 18:00	John Chen	xin.chen9@student.unsw.edu.au

9 Plagiarism policy

Your program must be entirely your own work. Plagiarism detection software might be used to compare submissions pairwise (including submissions for any similar projects from previous years) and serious penalties will be applied, particularly in the case of repeat offences.

Do not copy from others. Do not allow anyone to see your code. Please refer to the UNSW Policy on Academic Honesty and Plagiarism if you require further clarification on this matter.