UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA
Denkleiers • Leading Minds • Dikgopolo tša Dihlalefi

COS711 Assignment 2

Using Neural Networks to Predict CO2 Emissions

Due date: 25 September 2023, at 23h30

1 General instructions

You have to submit a pdf document, containing a technical report wherein you describe what you have done, present and discuss your findings. Guidelines for writing the report are provided in this specification document. The report will be checked for plagiarism using the Turnitin system, and should be submitted through the ClickUp system. You are advised but not required to typeset your report in LaTeX.

2 Greenhouse gas emissions

Greenhouse gas (GHG) emissions from human activities strengthen the greenhouse effect, contributing to climate change. Carbon dioxide (CO2), from burning fossil fuels such as coal, oil, and natural gas, is one of the most important factors in causing climate change.

Quoting a United Nations report, *"Climate change is likely to increase the prevalence of vector-borne diseases, such as malaria and dengue fever, and may increase the intensity of severe weather events. It is likely to lead to an increase in water levels and serious flooding, and at the same time cause water scarcity in arid regions. Climate change is expected to irreversibly damage some natural resources and ecosystems. Overall, climate change is projected to deliver a devastating combination of adverse impacts for the world's poor, both because of geography and low income, making adaptation to climate change much more difficult. While developing countries have contributed the least to the problem, they are expected to bear the brunt of the impact of climate change, which threatens to jeopardize many of the developmental gains that have already been achieved."*

3 Supervised Learning

For this assignment, you will work with a dataset that records **sustainable energy indicators** and other climate change related factors across 176 countries from 2000 to 2020. Specific column of interest is CO2 emissions: your goal will be to construct a neural network (NN) model that can predict CO2 emissions based on the various indicators provided. Once the NN model has been trained, you will use it to forecast CO2 emissions for the next 5 years.

3.1 Data set

Download the data set from ClickUP. The data set was borrowed from Kaggle. The data is provided in a csv format, where each row represents a data set entry, and each column represents a data attribute. You will see that for each country, 20 rows of data are recorded, corresponding to the 20 years from 2000 to 2020. The 14th column, titled "`Value_co2_emissions_kt_by_country`", represents the target value that your NN will need to predict. As such, your task is to construct a NN for a *regression* task.

3.2 Your task

Your task is to train a NN to correctly predict the CO2 emissions based on the year and other indicators, as recorded in the data set. For the purpose of this assignment, you will need to perform the following steps: pre-process the data; optimise NN hyperparameters; perform CO2 emission predictions for 5 years for South Africa and two other countries of your choice.

3.2.1 Data preparation

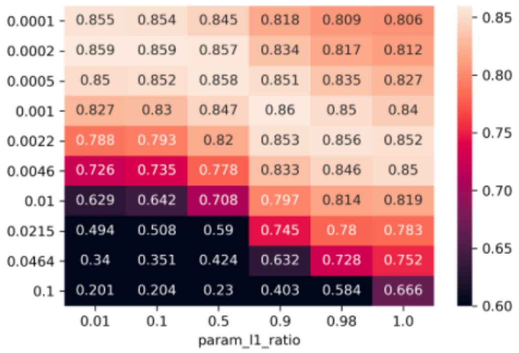
The given data contains numeric as well as categorical attributes that lie in various ranges. Analyse the data set, and pre-process it in a way that will make it possible for a NN to effectively discover the hidden relationships between inputs and outputs. Give extra thought to the following data set properties:

1. Country name is provided, as well as longitude and latitude per country. Are these variables important? Do you need all three? Which ones are likely to be the most informative?
2. The dataset contains missing values. Should columns/rows with missing values be excluded, or can the missing values be imputed in a meaningful way?
3. Some columns, such as "Electricity from nuclear (TWh)", contain predominantly zeros. What is the implication of this, how should such columns be treated?

3.2.2 Hyperparameter Optimisation

As discussed in class, the performance of your NN model greatly depends on various hyperparameters, such as the NN architecture, activation functions, error (loss) function, optimisation algorithm choice, optimisation algorithm parameters, etc. You will have to choose the hyperparameter values for your NN model. Your report **must** contain a section justifying all hyperparameter choices. Two justifications are acceptable: (1) theoretical insight; (2) empirical evidence. I.e., if you cannot decide on a value for a certain hyperparameter analytically, you have to run some experiments to see which value performs better than others.

You must empirically compare at least two different hyperparameters of your choice. While you are welcome to thoroughly optimise more that just two hyperparameters, it is required of you to perform a **grid search** over any two hyperparameters of your choice. Grid search implies selecting a set of values for each hyperparameter, and evaluating the model for every combination of values between the two sets. Visualise the results of the grid search using a heatmap such as the one shown below, where x-axis represents the values of hyperparameter A, y-axis represents the values of hyperparameter B, and each cell is colourised according to the model performance for a combination of (A,B):



Additionally, you must compare your NN to a simple perceptron, i.e. a NN without a hidden layer. It is a good sanity check that will give you an idea about your data's linearity.

NB: to compare the performance of any two hyperparameter values, you must obtain average performance for each hyperparameter across a few independent runs, as well as standard deviation. Is there statistical evidence that one value is more performant than another value? Consider using appropriate hypothesis testing to make an informed conclusion. Hint: if standard deviations do not overlap, you can safely conclude that the difference in performance is significant.

Note: since we are working with a regression task, you will not be able to measure the accuracy of your model in terms of the easily interpretable classification error. Read up on regression: how can regression models be compared to one another? Consider using metrics such as R-squared. As a sanity check, look at the absolute difference between the target outcomes and the predicted outcomes: how far off is your NN? You can also plot the actual/predicted CO2 emissions over the 20 years for a given country, and observe how well the curves correspond.

3.2.3 Predicting CO2 emissions for the next 5 years

Now that you have a model that is both optimised for performance and trained, you must use it to perform inference on CO2 emissions for the next five years. Use the last data entry for South Africa, and feed it to the NN, altering the year from 2021 to 2025. Pick two more countries from the list and repeat the process. Show the five-year prognosis on a plot and discuss it: does it look like your model is good at predicting the future? Is the forecast pessimistic, optimistic, or nonsensical? Try altering other attributes, such as the use of fossil fuels. Does it significantly alter the predictions? Does it seem that some attributes have a stronger effect on the output than others?

Discuss and interpret all your results thoroughly. Remember that simply pasting a table with numbers in it, or a graph with no explanation, will not yield any marks. Visualise the experimental data where appropriate: it is much easier to analyse your results when you can see them plotted next to one another in different colours. If you see that one approach is doing better than another, provide a hypothesis for why it is the case. Running the experiments is only half of the research process, the other and more important half is interpretation. Aim to derive as many insights from your results as you can.

4 Notes

- Implementation
  - You may use any programming language and platform.
  - You may use a machine learning/neural network library/framework.
- Report
  - You must report on all data preparation steps taken.
  - You must report on all algorithm hyperparameters used, and substantiate your choices.
  - Training and generalisation errors have to be reported. Remember to report means with the corresponding standard deviations, and report how many independent runs have been performed.

5 Marking and general guidelines

For this assignment you have to submit a **research report** where you discuss your findings. Your reports must follow the IEEE conference format (http://www.ieee.org/conferences_events/

conferences/publishing/templates.html). You may use the Latex or the Word template, however, it will serve as good academic writing practice to utilise LaTeX. There is also a strict page limit of **8 pages** for this assignment. Given the imposed two column format it would require a substantial amount of writing to exceed this limit.

This is not a course in technical and report writing; however, you should at least attempt to follow some accepted document writing techniques and make your report as readable as possible. You are more likely to obtain a higher mark if your report generates a good impression with the marker and is void of general errors like spelling and grammar mistakes.

A typical research report would consist of the following sections:

1. **Abstract**

   The abstract should briefly summarise the purpose and findings of the report.

2. **Introduction**

   The introduction sets the stage for the remainder of your report. You usually have very general statements here. The introduction prepares the reader for what to expect from reading your report. In general, the introduction should either contain or be a summary of your ENTIRE report. Keep the introduction concise, try to limit it to 1 page maximum.

3. **Background**

   A very high level discussion on the problem domain and the algorithms and/or approaches that you have used. This section is typically where the "base cases" of concepts that appear throughout the remainder of your report are discussed. It is also an ideal place to refer a reader to other sources containing relevant information on the topic which is outside the scope of your assignment. Remember to discuss very generally. After reading this section the marker should be able to determine whether or not you understand the techniques that you are using. Try to limit this section to 1 page maximum.

4. **Experimental Set-Up**

   In this section you discuss how you approached, implemented and solved your assignment. Mention the values set for the algorithm's hyperparameters, how many simulations you have run and what the characteristics for candidate solutions to your problems are. After reading this section (in addition to the background) the reader should be able to duplicate your experiments to obtain similar results to those obtained by you. This is also the section where your discussion specialises on the concepts mentioned in the background section. Be very specific in your discussions in this section.

5. **Research Results**

   This is the section where you report your results obtained from running the experiments as discussed in the experimental set-up section. You have to give, at the very least, the averages and the standard deviations for all the experiments/simulations. Graphing your results is advisable, and no conclusions regarding the superiority of one approach over another can be made without some form of statistical reasoning. Training, and generalisation errors have to be reported. Thoroughly discuss the results that you have obtained, and reason about why you obtained the results that you have. Answer questions like "are these results to be expected?" and "why these results occurred?" and "would different circumstances lead to different results?"

6. **Conclusion(s)**

   Very general conclusions about the assignment that you have done. This section "answers" the questions and issues that you have raised and investigated. This section is, in general, a summary of what you have done, what the results were, and finally what you concluded from these results. This is the final section in your document, so be sure that all the issues raised up until now are answered here. This is also the perfect section to discuss what you have learnt in doing this assignment.

Please **do not** include any code or pseudocode in the report. Research reports must focus on the scientific contributions. We assume that you are proficient at coding – you do not have to prove it anymore!

5.1 Marking

The following general breakdown will be used during the assessment of this assignment:

Category	**Mark Allocation**
Report Structure	5 marks
Background	5 marks
Data Preparation	20 marks
Experimental Setup	10 marks
Hyperparameter Optimisation	25 marks
Future CO2 Emissions Prediction	25 marks
Conclusions	5 marks
References	5 marks
TOTAL	100 marks

Submit only the PDF report. No additional files of any sort should be submitted, although the lecturer may request to inspect your code in case your report raises questions. Upload the PDF file to the appropriate assignment slot on ClickUp. Multiple uploads are allowed, but only the last one will be marked. The deadline is **25 September 2023, at 23h30** .