BUSS6002 Assignment

Semester 2, 2024

Instructions

- Due: at 23:59 on Friday, October 25, 2024 (end of week 12).
- You must submit a **written report** (in PDF) with the following filename format, replacing `STUDENTID` with your own student ID: `BUSS6002_STUDENTID.pdf`.
- You must also submit a **Jupyter Notebook** (`.ipynb`) file with the following filename format, replacing `STUDENTID` with your own student ID: `BUSS6002_STUDENTID.ipynb`.
- There is a limit of 6 A4-pages for your report (including equations, tables, and captions).
- Your report should have an appropriate title (of your own choice).
- Do not include a cover page.
- All plots, computational tasks, and results must be completed using Python.
- Each section of your report must be clearly labelled with a heading.
- Do not include any Python code as part of your report.
- All figures must be appropriately sized and have readable axis labels and legends (where applicable).
- The submitted `.ipynb` file must contain all the code used in the development of your report.
- The submitted `.ipynb` file must be free of any errors, and the results must be reproducible.
- You may submit multiple times but only your last submission will be marked.
- A late penalty applies if you submit your assignment late without a successful special consideration. See the Unit Outline for more details.
- Generative AI tools (such as ChatGPT) may be used for this assignment but you must add a statement at the end of your report specifying how generative AI was used. E.g., *Generative AI was used only used for editing the final report text.*
- Hint! It is highly recommended that you finish the week 10 tutorial before starting this assignment.

Description

One of the UN Sustainable Development Goals is 'climate action' (goal 13). In this assignment, you are conducting a study that compares the predictive performance between three families of basis functions: *polynomial*, *piece-wise constant*, and *piece-wise linear*, for a linear basis function (LBF) model designed to predict the global *surface air temperature.* The aim is to investigate which family of basis functions is most suited for modelling the relationship between time and temperature.

You are provided with the ERA5 surface air temperature dataset, which is widely used in climate research, weather forecasting, and environmental monitoring. The dataset contains 1,017 observations of monthly surface air temperature in degrees Celsius (`temp`) from January 1940 to September 2024. It also contains the year (`year`) and month (`month`) for which the temperature is observed. A scatter plot of the dataset is shown in Figure 1.

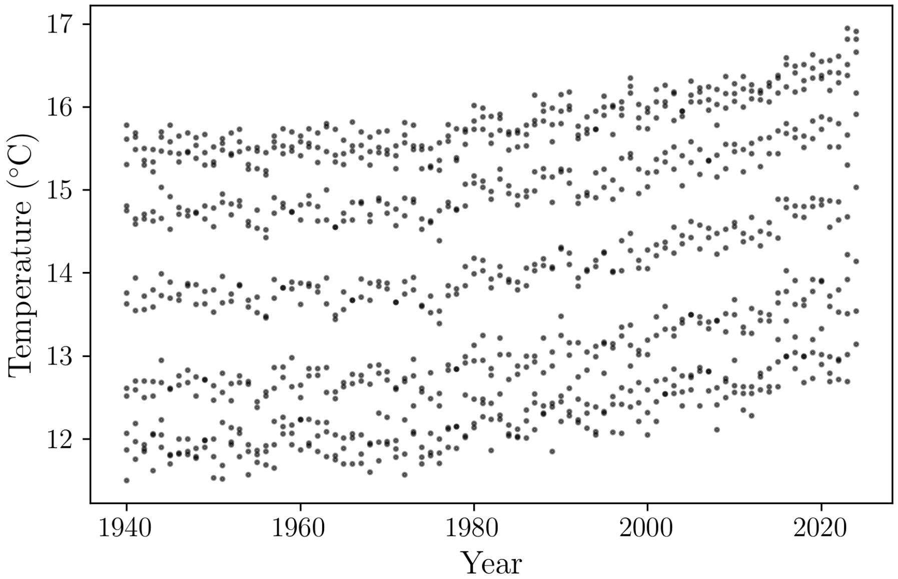

Figure 1: ERA5 surface air temperature from January 1940 to September 2024.

The specific LBF model being considered in your study is given by

$$y = \mathbf{u}^\top \boldsymbol{\alpha} + \boldsymbol{\phi}(x)^\top \boldsymbol{\beta} + \varepsilon,$$

where y is the surface air temperature, x is year, and ε is a random noise; $\mathbf{u} := [u_2, \dots, u_{12}]^\top$ is a binary vector of dummy variables, with $u_i = 1$ if y is observed in month i and $u = 0$ otherwise; $\boldsymbol{\phi}(x)$ denotes the vector of basis function values; the parameter vectors are $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Three families of basis functions are considered for computing $\boldsymbol{\phi}(x)$; the first family is the set of polynomial basis functions $\boldsymbol{\phi}(x) := [1, \phi_1(x), \dots, \phi_p(x)]^\top$, with

$$\phi_i(x) := x^i.$$

The second family is the set of piece-wise constant basis functions $\boldsymbol{\phi}(x) := [1, \gamma_1(x), \dots, \gamma_k(x)]^\top$, with

$$\gamma_i(x) := I(x > t_i),$$

where $I(x > t_i)$ is an *indicator function* defined by

$$I(x > t_i) := \begin{cases} 1 & \text{if } x > t_i \\ 0 & \text{if } x \leq t_i. \end{cases}$$

The break points $\{t_i\}_{i=1}^{k}$ are calculated according to

$$t_i := x_{\min} + \frac{i(x_{\max} - x_{\min})}{k+1}, \tag{1}$$

where $x_{\min}$ and $x_{\max}$ denote the smallest and largest observed values of x, respectively. The third family is the set of piece-wise linear basis functions $\boldsymbol{\phi}(x) := [1, x, \lambda_1(x), \ldots, \lambda_k(x)]^\top$, with

$$\lambda_i(x) := (x - t_i)I(x > t_i),$$

where t_i is given by Equation (1).

Before comparing the three basis function families, you must set the degree p for the polynomial model, and the number of break points k for the piece-wise constant and piece-wise linear models. The hyperparameter value for each basis function family should be selected using a validation set, by minimising the validation mean squared error (MSE).

For the polynomial model, the optimal value of p should be selected by exhaustively searching through an equally-spaced grid from 1 to 10, with a spacing of 1:

$$\mathcal{P} := \{1, 2, \ldots, 10\}.$$

For the two piece-wise models, you should select the optimal values of k by exhaustively searching through another equally-spaced grid from 1 to 30, with a spacing of 1:

$$\mathcal{K} := \{1, 2, \ldots, 30\}.$$

Once the optimal values of the hyperparameters are chosen for all basis function families, you will be able to compare the predictive performance between the three using a test set (i.e., by comparing the test MSE between the three optimally selected models).

Report Structure

Your report must contain the following four sections:

Report Title

1 Introduction (0.5 pages)

- Provide a brief project background so that the reader of your report can understand the general problem that you are solving.
- Motivate your research question.
- State the aim of your project.
- Provide a short summary of each of the rest of the sections in your report (e.g., "The report proceeds as follows: Section 2 presents . . . ").

2 Methodology (2 pages)

- Define and describe the LBF model.
- Define and describe the three choices of basis function families being investigated.
- Describe how the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated given the hyperparameter value. Discuss any potential numerical issues associated with the estimation procedure.
- Describe how the hyperparameter value can be determined automatically from data (as opposed to manually setting the hyperparameter to an arbitrary value).
- Describe how the performance of the three families of basis functions is compared given the optimal hyperparameter value.

3 Empirical Study (2.5 pages)

- Describe the datasets used in your study.
- Present (in a table) the selected hyperparameter value for each basis function family.
- Describe and discuss the table of selected hyperparameters.
- Visually present (using plots) the predicted response values for each basis function family in the test set.
- Describe and discuss the plots of predicted values.
- Present (in a table) the test MSE values for each basis function family.
- Describe and discuss the table of test MSE values.
- Report the temperature forecasts for October, November, and December of 2024 given by the model with the smallest test MSE. Include a brief description of how these forecasts are obtained.

4 Conclusion (0.5 pages)

- Discuss your overall findings / insights.
- Discuss any limitations of your study.
- Suggest potential directions of extending your study.

Rubric

This assignment is worth 30% of the unit's marks. The assessment is designed to test your computational skills in implementing algorithms and conducting empirical experiments, as well as your communication skills in writing a concise and coherent report presenting your approach and results. The mark allocation across assessment items is given in Table 1.

Assessment Item	Goal	Marks
Section 1	Introduction	4
Section 2	Methodology	10
Section 3	Empirical Study	16
Section 4	Conclusion	3
Overall Presentation	Clear, concise, coherent, and correct	5
Jupyter Notebook	Reproducable results	2
Total		40

Table 1: Assessment Items and Mark Allocation