ML Programming assignment IV

Name：林欣妤
Student number：314652023

September 30, 2025

1 Description of the Problem

In this assignment, our first step is to carefully preprocess the raw data in order to generate two separate datasets: one for classification and another for regression. After obtaining these clean and well-structured datasets, we proceed to train the two models using different approaches tailored to their respective tasks. Specifically, for the classification model, our goal is to predict the probability that the temperature at any given latitude and longitude is valid, effectively distinguishing between usable and missing or erroneous data points. In contrast, for the regression model, we aim to predict the actual temperature at the specified location, based solely on the input latitude and longitude, using only the valid temperature measurements. This separation of tasks ensures that each model can focus on learning the most relevant patterns for its intended purpose.

2 Programming process

The program is divided into three parts: preprocessing, the classification model, and the regression model. In addition, for the regression model, I found during this assignment that the previous use of the validation set might not have been very precise, so I made corrections to its usage in this assignment.

2.1 Preprocessing

First, in preprocessing, we begin by carefully reading the data from the file **O-A0038-003.xml**, extracting not only the initial latitude and longitude values but also all the temperature measurements recorded in the dataset. By utilizing the longitude and latitude resolutions provided in the XML file (0.03 degrees for both), we are able to compute the precise geographical coordinates corresponding to each temperature reading, ensuring that all spatial information is accurately aligned. For the classification dataset, any temperature values equal to -999 are considered invalid and are therefore labeled as 0 to indicate missing or unusable data. In the case of the regression dataset, we remove all invalid temperature entries entirely, retaining only the valid measurements so that the model can be trained on reliable and meaningful data. This careful preprocessing ensures that both datasets are clean, properly structured, and ready for subsequent modeling tasks.

- Input: O-A0038-003.xml
- Output: regression_dataset.csv, classification_dataset.csv

2.2 The Classification Model

For the classification model, since the decision boundary is nonlinear, using logistic regression (which finds a single line to separate the data into two groups) is difficult. Therefore, we considered transforming the data into a higher-dimensional space and using a hyperplane to separate the data, following the classification concept. We used **SVM** for training to find the classification boundary. The training data was split into five folds for training and validation to determine the best SVM model parameters. Additionally, Platt scaling was applied to estimate the probability that each point is valid (i.e., the probability that label = 1) [2, 3].

The figure below shows our results, displaying the probability distribution that each latitude-longitude temperature value is valid. Colors indicate the probability: the higher the probability, the redder the color; the lower the probability, the bluer the color (i.e., the higher the probability that label = 0). In the figure, it can be seen that the training results are less accurate near the main island ' s boundaries and the outlying islands.

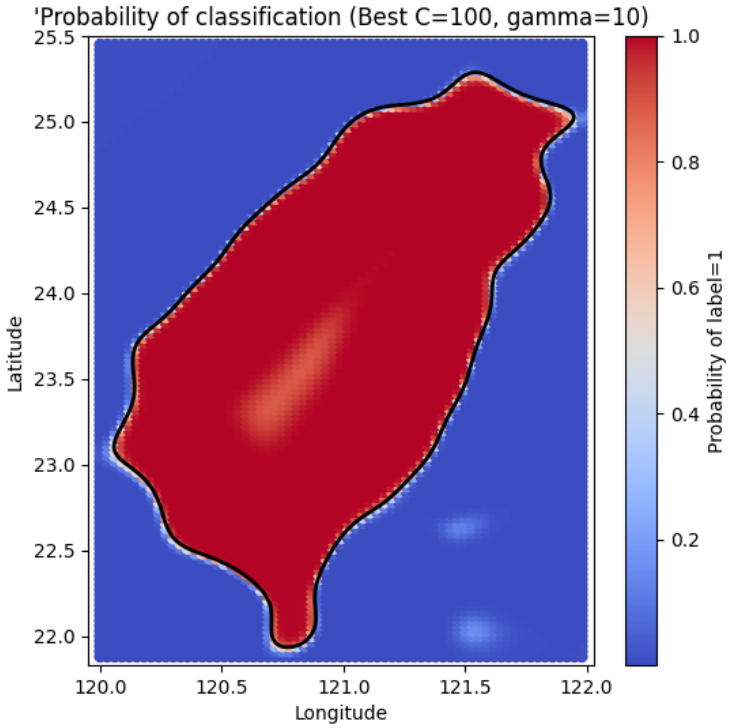

Figure 1: the probability that label = 1

2.3 The Regression Model

The regression model is generally similar to previous neural network approaches. However, during the first training run, I obtained unusual results, possibly because I was using ReLU activation. After investigation, it seems that data standardization can resolve this issue. The final results are shown in the figure 2 the left plot shows the temperatures predicted by the model, the right plot shows the original temperature distribution, and the Figure 3 shows the loss curve of our model (using the standardized data).

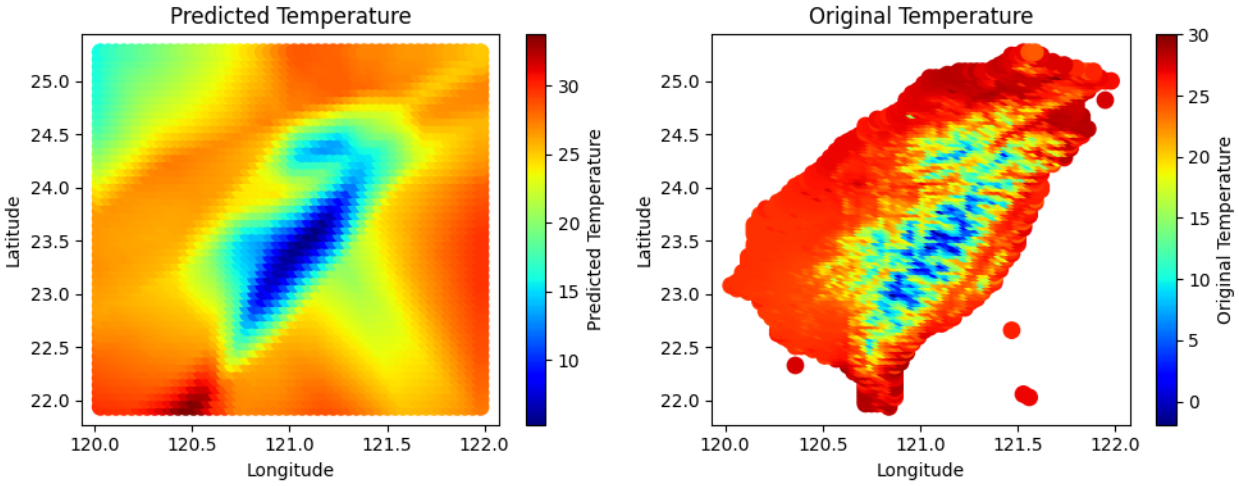

Figure 2: Temperature

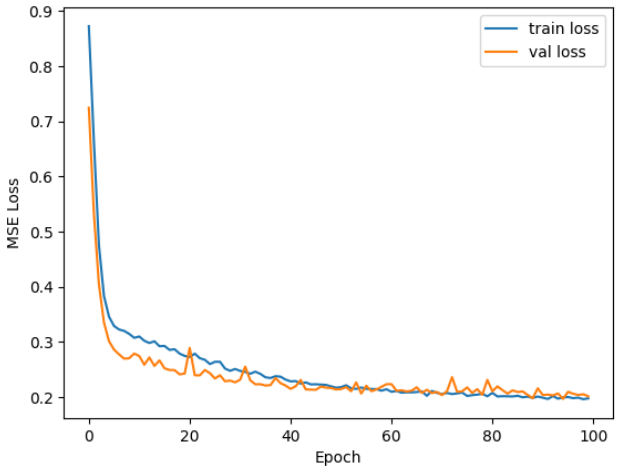

Figure 3: Loss

References

[1] Chat-GPT(Apply GPT to revise and correct the English content of the report, and ask about some programming techniques)

[2] Python 機器學習-多元分類的 5 種模型
https://medium.com/@imirene/Python 機器學習-多元分類的 5 種模型-f7b6026c5ce6

[3] 白話文講解支持向量機 (二) 非線性 SVM
https://notes.andywu.tw/2020/白話文講解支持向量機二-非線性 svm/