Report

Part I — Gaussian Discriminant Analysis (GDA)

1. Theoretical Background

Gaussian Discriminant Analysis (GDA) is a **probabilistic generative model** for supervised classification.
It assumes that data from each class follow a multivariate normal distribution, each characterized by its own mean vector μ_{k} and covariance matrix Σ_{k}.
For a given observation x, the conditional probability density function for class k is:

$$\mathrm{p}(\mathrm{x}|\mathrm{y}=\mathrm{k}) \ = \frac{1}{(2\pi)^{\mathrm{d}/2}|\Sigma_{\mathrm{k}}|^{1/2}}\mathrm{e}^{-\frac{1}{2}(\mathrm{x}-\mu_{\mathrm{k}})^{\mathrm{T}}\Sigma_{\mathrm{k}}^{-1}(\mathrm{x}-\mu_{\mathrm{k}})}$$

Using Bayes' theorem, the posterior probability of each class is computed as:

$$\mathrm{p}(\mathrm{y}=\mathrm{k}|\mathrm{x}) \ = \frac{\mathrm{p}(\mathrm{x}|\mathrm{y}=\mathrm{k})\mathrm{p}(\mathrm{y}=\mathrm{k})}{\Sigma_{\mathrm{j}}\mathrm{p}(\mathrm{x}|\mathrm{y}=\mathrm{j})\mathrm{p}(\mathrm{y}=\mathrm{j})}$$

The prediction rule is then to assign the class label k with the higher posterior probability.
When the covariance matrices of the two classes are identical ($\Sigma_0 = \Sigma_1$), the decision boundary is linear — this special case is known as Linear Discriminant Analysis (LDA).
When the covariance matrices differ ($\Sigma_0 \neq \Sigma_1$), the decision boundary becomes **quadratic**, which defines **Quadratic Discriminant Analysis (QDA)**.

In the given temperature dataset, each record represents a geographical grid point with longitude and latitude as features, and a binary label indicating whether the temperature measurement is valid or missing.
Since these valid and invalid regions are spatially clustered and approximately Gaussian in structure, GDA and QDA are appropriate models for classification.

2. Model Training and Evaluation

The QDA model was trained on the full dataset, where:

- Class 1 represents valid temperature measurements.
- Class 0 represents invalid or missing data (values ≤ −900).

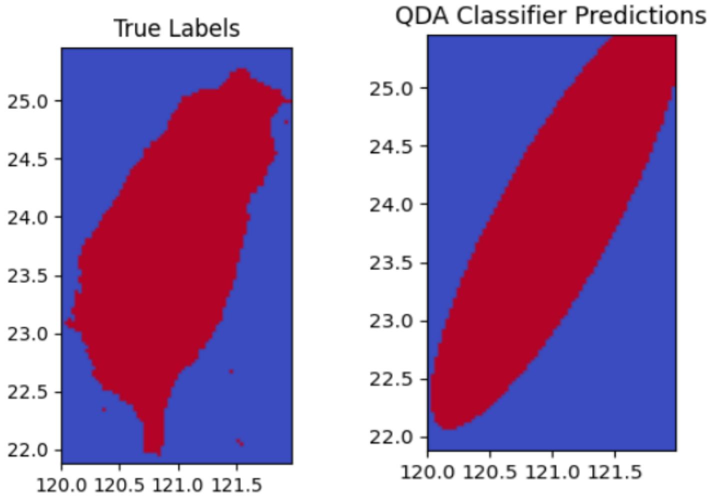
Since the dataset is not randomly sampled but spatially structured, training on the full grid provides a realistic evaluation of model performance.
The primary evaluation metric used was **classification accuracy**, defined as:

$$\text{Accuracy} = \frac{Correct\ predictions}{\text{Total number of samples}}$$

The trained QDA model achieved an overall accuracy of approximately **91%**, indicating that it correctly distinguishes valid and invalid grid cells in most regions. This performance confirms that the Gaussian assumption fits the data distribution reasonably well.

3. Visualization of the Decision Boundary

The learned QDA decision boundary exhibits a **nonlinear** separation between the two classes.
When visualized across the spatial grid, the QDA model effectively outlines the geographic region of valid temperature readings while excluding areas with missing or erroneous values.
The classification map shows that the model successfully captures the spatial transition between land-based and non-land or missing data regions, reflecting the expected real-world structure of the dataset.



Part II — Piecewise Regression Model

1. Model Concept and Motivation

While the QDA model provides class labels, it cannot directly predict temperature values.
To extend the analysis, we define a **piecewise regression function** that integrates both classification and regression.
The piecewise model combines two sub-models:

1. A **classifier C(x)** — a neural network that predicts whether a point contains valid data.
2. A **regressor R(x)** — another neural network trained only on valid data to predict temperature values.

These two components are combined into a single composite function:

$$\mathrm{h(x)} = \begin{cases} \mathrm{R(x)}, & \text{if } \mathrm{C(x)} = 1, \\ \text{-999}, & \text{otherwise} \end{cases}$$

This design ensures that the regression model only operates in regions deemed valid by the classifier,
while all other areas are automatically marked as missing (−999).
It provides a clean separation between “predictable” and “undefined” spatial zones.

2. Model Training and Application

Both neural networks were trained independently:

- The **classifier** was optimized using *cross-entropy loss* to distinguish valid and invalid locations.
- The **regressor** was optimized using *mean squared error (MSE)*, trained exclusively on valid samples.

After training, the two models were integrated through the piecewise definition.
The combined model h(x) was then applied to the entire dataset.
For each grid coordinate, the classifier first determined whether the point belonged

to a valid region.
If valid, the regressor predicted a temperature value; otherwise, the function output −999.

This process verified that the piecewise formulation behaves as expected:
the final output correctly produces continuous temperature estimates only in valid areas.

3. Experimental Results

The model performances are summarized below:

Model Component	**Description**	**Evaluation Metric**	**Result**
QDA Classifier	Quadratic probabilistic classifier	Accuracy	**0.91**
Neural Classifier	Learned nonlinear boundary	Accuracy	**≈0.94**
Neural Regressor	Temperature predictor (valid zones)	RMSE	**1.54°C**
Piecewise Function (h(x))	Combined classification-regression model	Behavior	**Correct (conditional output verified)**

These results show that both classification and regression components perform effectively and complement each other within the combined framework.

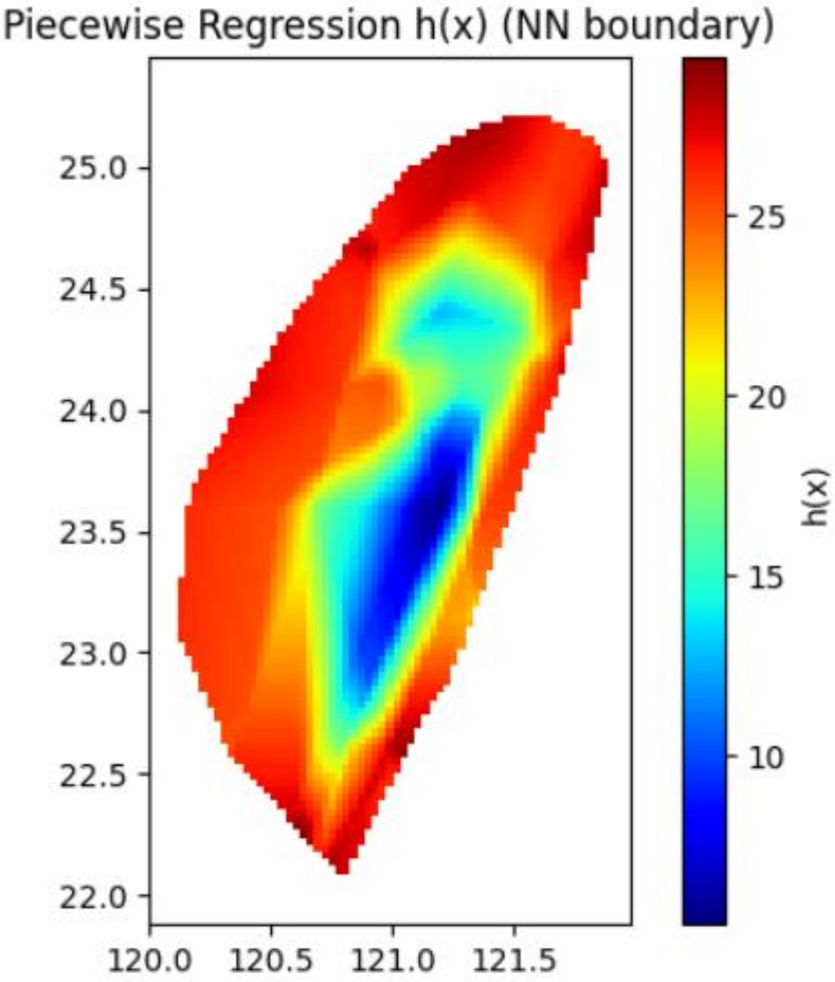
4. Visualization and Model Behavior

Two key visualizations illustrate the model’s performance:

1. **QDA Classification Map** — Demonstrates how QDA partitions the spatial domain into valid and invalid regions using a curved, nonlinear decision boundary.

2. **Piecewise Regression Map (h(x))** — Displays predicted temperature values across the grid, with color indicating temperature intensity and blank (white) regions representing invalid or missing data.
   The valid regions correspond precisely to the classifier's decision boundary, confirming that the piecewise function operates correctly.

The visual consistency between classification results and regression outputs provides clear evidence that the model structure is well-integrated and logically sound.


Piecewise Regression h(x) (NN boundary)
h(x)