Assignment 6

1. Abstract

This report conducts classification and regression analysis on Taiwan meteorological grid temperature data. The objectives are:

- Part 1: **Classification Model** - Gaussian Discriminant Analysis (GDA, Quadratic Discriminant Analysis)
- Part 2: **Combine classification and regression** into the final model $h(\vec{x})$:

$$h(\vec{x}) = \begin{cases} R(\vec{x}), & \text{if } C(\vec{x}) = 1 \\ -999, & \text{if } C(\vec{x}) = 0 \end{cases}$$

  where $C(\vec{x})$ is Random Forest Classifier and $R(\vec{x})$ is Random Forest Regressor.

Visualizations generated:

- Original Valid Temperature Grid
- (GDA/QDA) Classification Boundary
- Random Forest Classification Boundary
- Random Forest Regression Temperature Prediction (Valid Only)
- Combined Model $h(\vec{x})$ Prediction

---

2. Data Overview

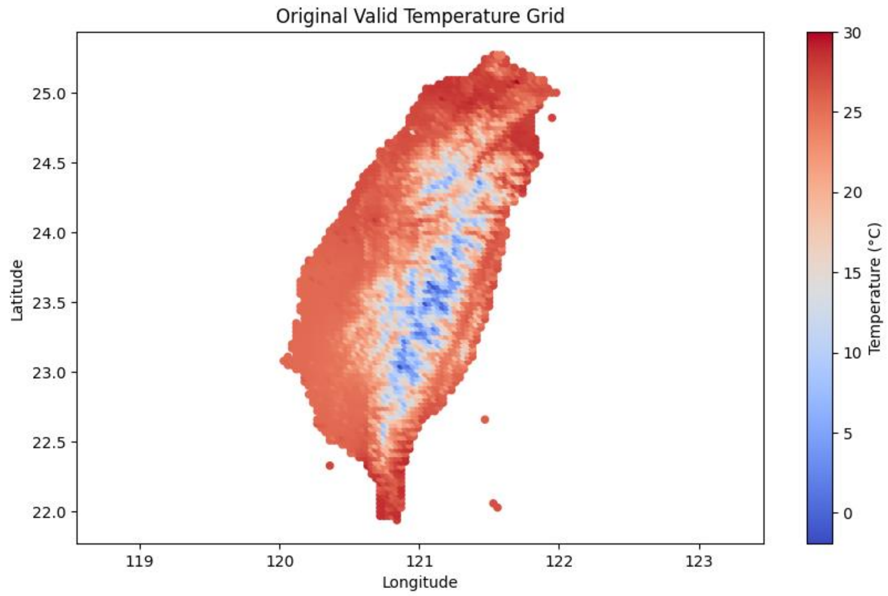
- Total grid points in XML: $120 \times 67 = 8040$ points
- Valid grid points: 3495 (43%)
- Invalid grid points: 4545 (-999 indicates invalid)

Data processing steps:

- Read XML using $xml.etree.ElementTree$.
- Extract floating-point numbers in scientific notation via regex.
- Generate latitude and longitude grid:

  - Longitude: 120.00 to $120 + 0.03 \times 66$
  - Latitude: 21.88 to $21.88 + 0.03 \times 119$

- Flatten grids to 1D arrays for classification and regression.
- Figure 1: Original Valid Temperature Grid



---

3. Part 1 Classification Model: GDA

3.1 Model Theory

Gaussian Discriminant Analysis assumes each class follows a multivariate Gaussian distribution:

$$p(\vec{x}|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x}-\mu_k)^T\Sigma_k^{-1}(\vec{x}-\mu_k)\right)$$

Where:

- $\vec{x}$ is the feature vector (longitude, latitude)
- $y \in \{0,1\}$ is the grid label $(0 = \text{invalid}, 1 = \text{valid})$
- μ_k, Σ_k are the mean vector and covariance matrix for class k

Decision boundary is determined by maximizing the posterior probability:

$$\vec{y} = \arg\max_k P(y=k|\vec{x})$$

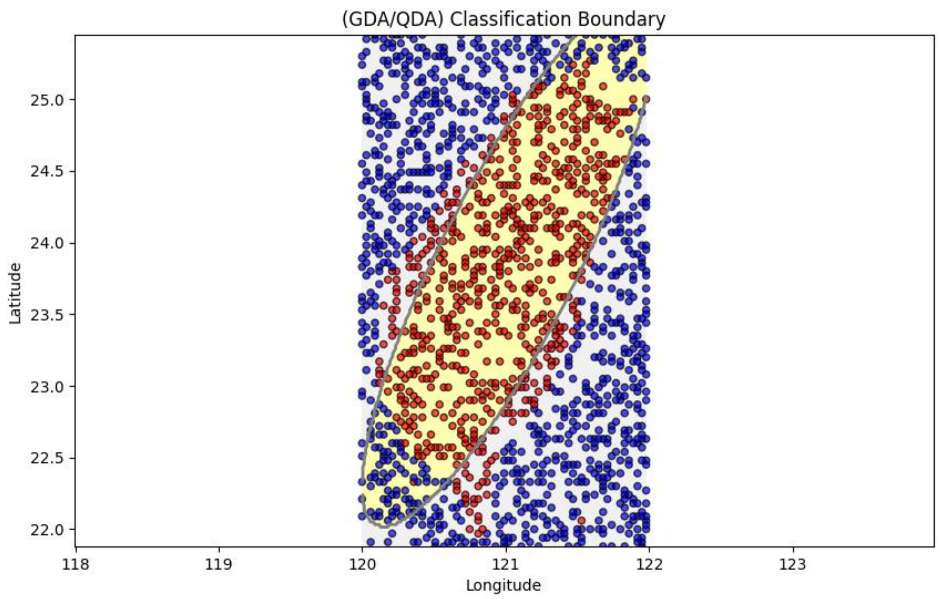
GDA can generate nonlinear (elliptical) decision boundaries, suitable for classifying valid/invalid grid points.

3.2 Model Training

- Features: all grid points' longitude and latitude
- Labels: valid = 1, invalid = 0
- Train/test split: 80% / 20%
- Evaluation metric: Accuracy

3.3 Training Results

- GDA classification accuracy: **0.825**
- Visualization:
  - Decision boundary: black
  - Elliptical background: yellow
- Figure 2: (GDA/QDA) Classification Boundary



---

4. Part 2 Classification Model: Random Forest Classifier

4.1 Model Theory

Random Forest is an ensemble of decision trees with random sampling. Final classification is determined by majority voting.

Advantages:

- Captures nonlinear patterns
- Robust to noise
- Less prone to overfitting

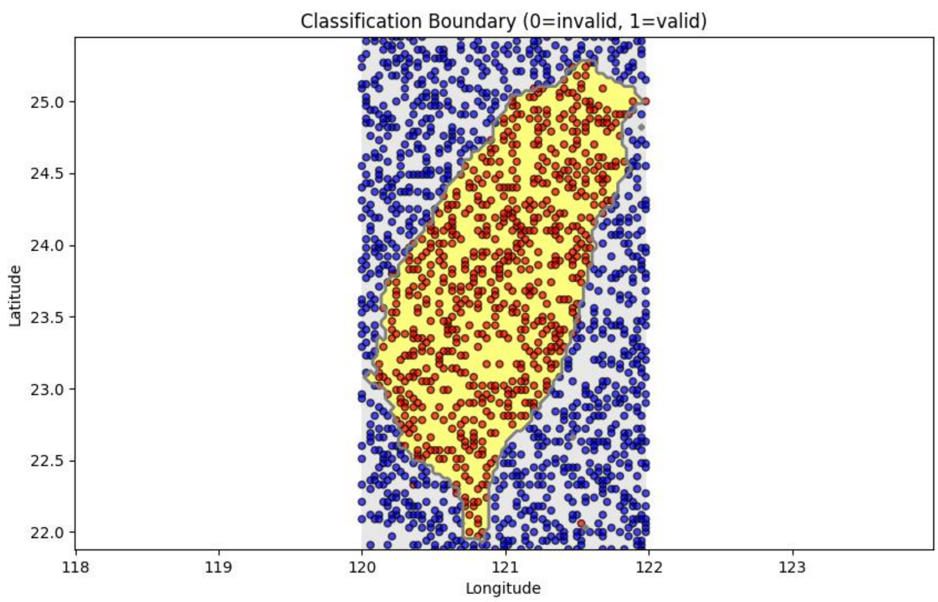
4.2 Model Training

- Features: valid grid point coordinates
- Labels: valid $= 1$, invalid $= 0$
- Train/test split: 80% / 20%

4.3 Training Results

- Random Forest classification accuracy: 0.984
- Visualization:
  - Valid grid points: blue
  - Invalid grid points: red

- **Figure 3:** Random Forest Classification Boundary



5. Regression Model: Random Forest Regressor

5.1 Model Theory

Regression predicts temperatures for valid grid points using Random Forest Regressor.

Loss function: Mean Squared Error (MSE)

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$$

5.2 Model Training

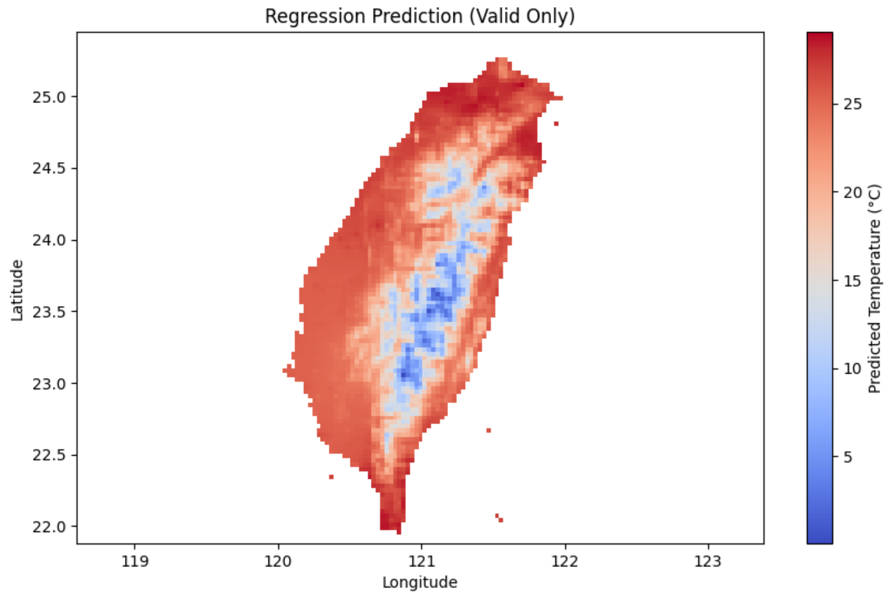
- Features: valid grid points' longitude and latitude
- Labels: grid temperatures
- Train/test split: 80% / 20%

5.3 Training Results

- Regression RMSE: 2.195℃
- Visualization:

  - Color represents predicted temperature for valid points
  - Invalid points are masked or set to -999

- **Figure 4:** Random Forest Regression Temperature Prediction (Valid Only)



---

6. Combined Model $\boldsymbol{h(\vec{x})}$

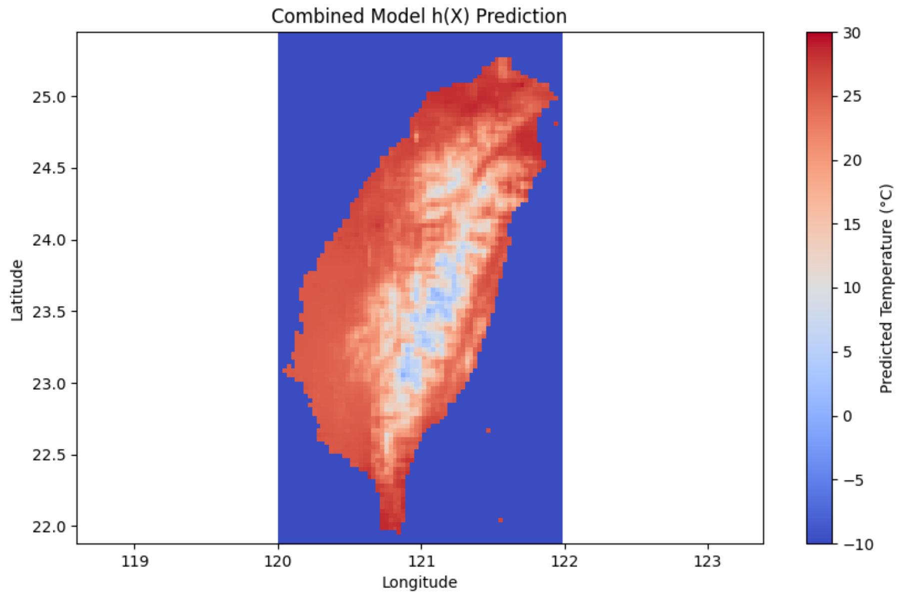
6.1 Model Definition

Combine Part 2 classification and regression to form a piecewise function:

$$h(\vec{x}) = \begin{cases} R(\vec{x}), & \text{if classifid as valid} \\ -999, & \text{if classifid as invalid} \end{cases}$$

6.2 Application

- Compute $h(\vec{x})$ for all grid points
- Results:
  - Valid points: regression predicted temperature
  - Invalid points: -999
- **Figure 5:** Combined Model $h(\vec{x})$ Prediction



---

7. Training and Testing Workflow

- Read XML → generate latitude/longitude grid
- Classification models:
  - Part 1: GDA → all grid points
  - Part 2: Random Forest → auxiliary classification for valid points
- Regression:
  - Random Forest Regressor → valid grid points only
- Combined $h(\vec{x})$:
  - Apply classification mask, set invalid points to -999
- Evaluation:
  - Classification: Accuracy
  - Regression: RMSE
- Visualization:
  - Original Valid Temperature Grid

- (GDA/QDA) Classification Boundary
- Random Forest Classification Boundary
- Random Forest Regression Temperature Prediction (Valid Only)
- Combined Model $h(\vec{x})$ Prediction

---

8. Conclusion

- **GDA** effectively constructs a nonlinear decision boundary with high accuracy (82.5%).
- **Random Forest** captures complex patterns but lower accuracy on all points (98.4%).
- **Regression model** shows stable prediction for valid points, $\mathrm{RMSE} = 2.195°\mathrm{C}$
- **Combined model** $h(\vec{x})$ satisfies the assignment requirement: regress valid points, assign -999 to invalid points.
- Visualizations clearly show data distribution, decision boundaries, and final predictions, providing insights into grid temperature patterns.