Machine Learning-Assignment 4

Wei-Ju Lin

October 1, 2025

1 Model and Description

1.1 Dataset

The raw files provide the geographical boundaries (longitude/latitude) and grid values, where some grid points are marked as `-999.0` to indicate missing or invalid values. The data was divided into two forms:

- **Classification dataset (classification.csv)**
  - Features: longitude (lon), latitude (lat).
  - Label: validity indicator (1 for valid, 0 for invalid).
  - Goal: predict whether a grid point contains valid meteorological data.
- **Regression dataset (regression.csv)**
  - Features: longitude (lon), latitude (lat).
  - Label: grid value (temperature in °C).
  - Goal: predict the continuous value at a given grid point.

1.2 Method

We apply **Radial Basis Function (RBF) feature mapping** to transform the two-dimensional geographic coordinates into a high-dimensional feature space and using logistic regression for classification and ridge regression for regression.

This approach allows the model to capture nonlinear spatial relationships, making it suitable for meteorological grid data.

2 Training Procedure

2.1 Data preprocessing

- Longitude and latitude were discretized at a resolution of 0.03°.
- The data matrix was flattened and exported as `classification.csv` and `regression.csv`.
- The classification dataset includes all grid points, while the regression dataset collects only valid values.

2.2 Feature mapping

- `RBFSampler` was used to transform (lon, lat) into 500-dimensional vectors.
- The gamma parameter controls the RBF scale:
  - Classification: $\gamma = 1.0$
  - Regression: $\gamma = 100.0$

2.3 Classification model

- Logistic Regression with `max_iter = 2000` to ensure convergence.
- Data split: 80% training, 20% testing, stratified by labels.
- Class weights balanced to address possible label imbalance.

2.4 Regression model

- Ridge Regression with L2 regularization to avoid overfitting.
- Data split: 80% training, 20% testing.

3 Results and Analysis

3.1 Data distribution

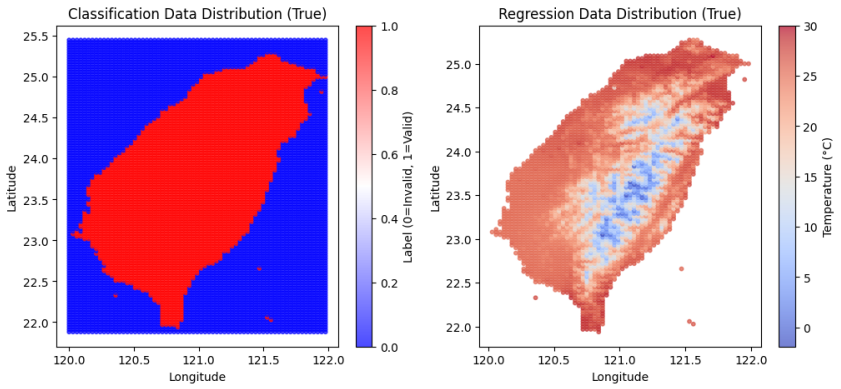

Figure 1: Data distribution

- The classification dataset shows missing values concentrated in certain regions.
- The regression dataset reveals spatially smooth variations in values.

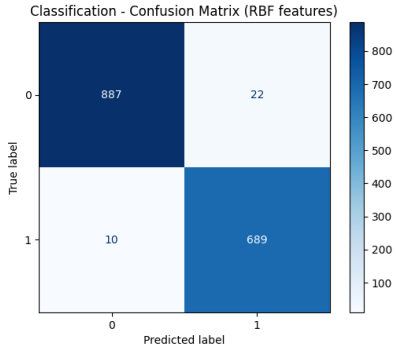

Figure 2: Confusion matrix

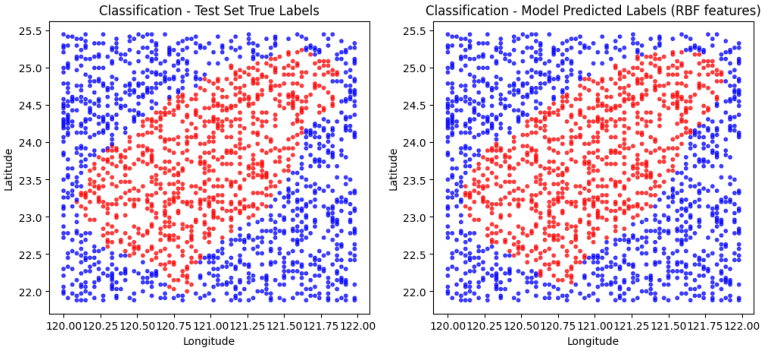

Figure 3: Classification results

3.2 Classification results

- Test accuracy: ≈ 0.97–0.99 (depending on random splits).
- Confusion matrix shows most valid points are correctly identified, with minor misclassifications at the boundaries.
- Visualizations demonstrate strong agreement between predicted and true labels.

3.3 Regression results

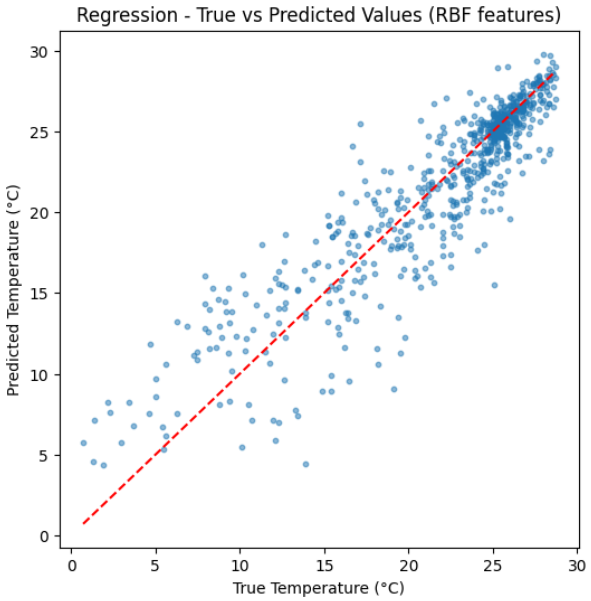

Figure 4: Regression results

- Test set performance: $R^2 \approx 0.820$, indicating that the model successfully captures the spatial patterns of temperature and provides reliable predictions, although a small portion of unexplained variability remains.
- Mean Squared Error (MSE): $\approx$ 6.092, the error is around 2.5°C, which means the model not only captures the overall distribution of temperature values but

also achieves a reasonably good level of precision, although some local variability remains unaccounted for.

- The plot shows that the model follows the overall trend. It works better at higher temperatures, but the errors are larger at lower temperatures.

4 Conclusion and Discussion

This study shows that using RBF features with simple linear models can work for spatial data classification and regression. The main findings are:

1. **Classification**: The model can tell valid and invalid grid points apart, which is helpful for removing missing data.
2. **Regression**: The model can fill in grid values, which is useful for making spatial prediction maps of meteorological variables.

However, there are several limits:

- We fixed the RBF feature size at 500. Bigger or harder datasets may need tuning or other kernel methods like Gaussian Process Regression.
- The model only used longitude and latitude. Adding more features like elevation or land type could make it more accurate.
- The value of γ in the RBF mapping was manually selected. A more systematic approach would be to include a parameter search procedure to automatically determine the optimal γ.

In short, the RBF with linear model gives a simple but effective baseline for meteorological grid data. Future work can add more features, use time-series data, and test higher-dimensional models.