1141ML-Week 6 Programming Assignment 02

Taiwan Weather Dataset Prediction - Function Package up

313652018 王宣瑋

September 2025

Tasks

Your task is to build a regression model that represents a piecewise smooth function. To do this, combine the two models from Assignment 4 into a single function.

(a) Implementation of the combined model

In this task, we combine two previously trained models into a single piecewise regression model $h(\mathbf{x})$. Let the classifier $C(\mathbf{x})$ determine whether a given sample belongs to a valid region, and let the regressor $R(\mathbf{x})$ predict a continuous value (temperature) in that region.

The combined model is defined as

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

Both C and R are implemented using `RandomForestClassifier` and `RandomForestRegressor`, respectively, ensuring a non-parametric and smooth approximation of the underlying data. The final function h is implemented in vectorized form to efficiently handle both single and batch predictions.

(b) Application and verification of the piecewise definition

The classifier was trained on the `classification_dataset.csv`, which labels land (class 1) and sea (class 0) regions based on latitude and longitude. The regressor was trained on the `regression_dataset.csv`, which provides temperature readings at corresponding coordinates. After training, $h(\mathbf{x})$ was applied to the regression dataset to verify that:

- regions with $C(\mathbf{x}) = 1$ produce continuous temperature values through $R(\mathbf{x})$, and
- regions with $C(\mathbf{x}) = 0$ are assigned a fixed value of -999.

(c) Explanation of how the combined function works

The combined model $h(\mathbf{x})$ first evaluates each input point with the classifier C. If the point is classified as valid ($C(\mathbf{x}) = 1$), the regressor R estimates a temperature value. Otherwise, $h(\mathbf{x})$ outputs -999, representing an invalid or out-of-region point. This construction ensures that h is continuous and smooth within valid areas, but piecewise-defined across the whole domain.

(d) Results and discussion

Classification performance:

The Random Forest classifier achieved an accuracy of **98.7%** on the test set, with only 26 misclassified samples out of approximately 2000. The confusion matrix is given by:

$$\begin{bmatrix} 1129 & 7 \\ 19 & 855 \end{bmatrix}$$

indicating very few false positives and false negatives. Spatially, the predicted classification map closely matches the ground truth, capturing the shape of the Taiwan landmass accurately.

Regression performance (within valid regions):

For samples classified as valid ($C(\mathbf{x}) = 1$), the regression model achieved:

$$\text{MAE} = 1.75^{\circ}\text{C}, \quad \text{RMSE} = 2.60^{\circ}\text{C}.$$

The scatter plot of predicted versus true temperature shows a strong alignment along the diagonal, demonstrating accurate predictions in most regions.

Piecewise model behavior:

The spatial map of $h(\mathbf{x})$ shows yellow points corresponding to valid regions (where $C = 1$ and $R(x)$ is used) and blue points for invalid regions (where $C = 0$ and $h(x) = -999$). Only a few isolated blue points appear near coastal boundaries, confirming that the classification step correctly identifies the valid land area. An error map for valid regions further reveals that larger prediction errors occur near edges or sparsely sampled regions, while most of the domain maintains low prediction error.

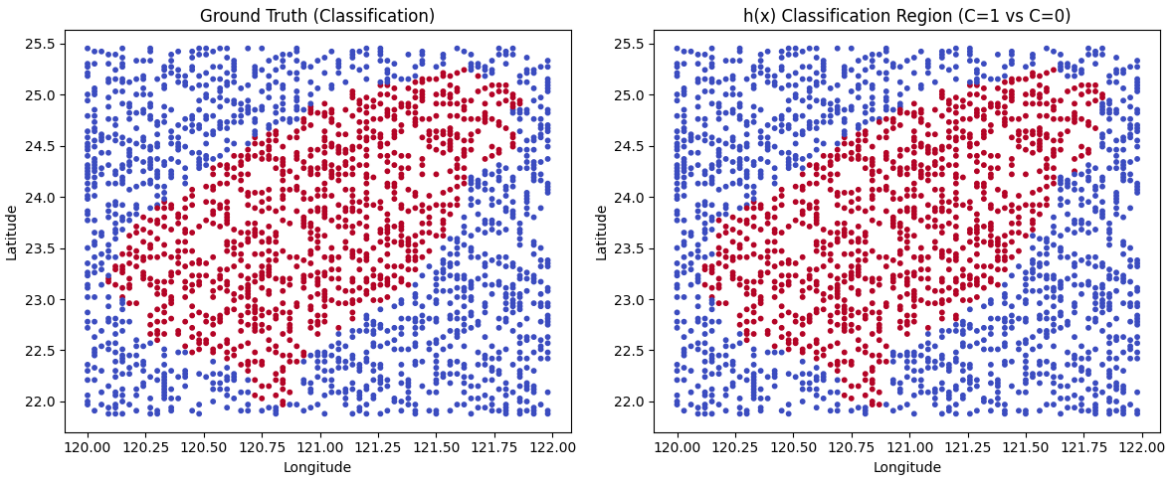
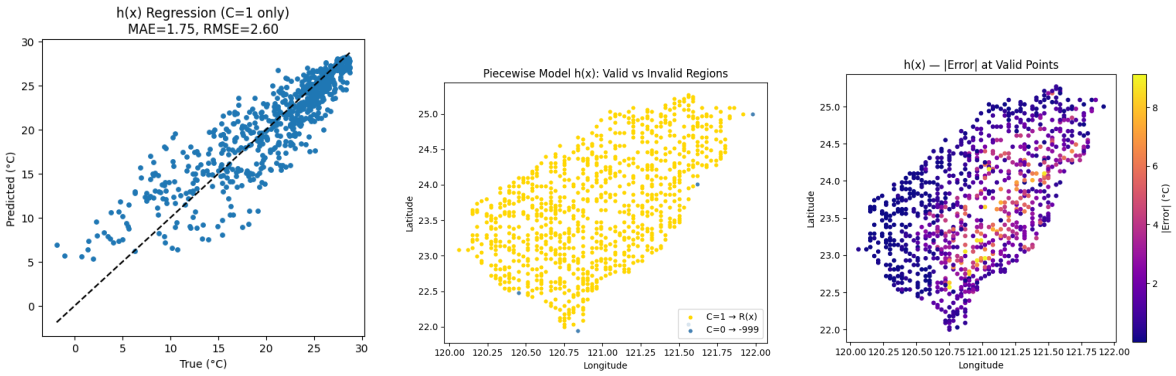


Figure 1: Classification comparison. Left: ground truth land/sea labels. Right: region where the piecewise model is active ($C(\mathbf{x})$=1). The Random Forest classifier reproduces the coastline closely, with only few off—boundary errors.

Discussion for the 3 pictures of Figure 2

(a) Regression accuracy: The scatter plot of predicted versus true temperatures shows that most points align closely with the $y = x$ line, indicating that the Random Forest regressor performs well in the valid region (C=1). The model achieves a mean absolute error (MAE) of approximately 1.75°C and a root mean square error (RMSE) of 2.60°C, suggesting that temperature predictions are generally accurate within the classified land area.



(a) $h(\mathbf{x})$ regression on valid region (C=1): True vs. Predicted. MAE = 1.75°C, RMSE = 2.60°C.

(b) Piecewise regions from $h(\mathbf{x})$: gold = C=1 (use R), blue = C=0 (output −999).

(c) Spatial $|y - \hat{y}|$ on valid points (C=1). Errors remain small over most areas and increase near edges/ sparser zones.

Figure 2: Piecewise model $h(\mathbf{x})$ results: (a) accuracy of regression on valid region, (b) valid vs. invalid spatial partition, and (c) spatial distribution of absolute errors.

(b) Piecewise structure: The spatial map of valid versus invalid regions confirms the expected piecewise definition. Yellow points represent locations where the classifier predicted C=1, and thus $R(x)$ provides a continuous temperature estimate. Blue points correspond to C=0, where $h(x)$ outputs the fixed value −999. The valid region clearly coincides with the Taiwan landmass, showing that $h(x)$ correctly restricts regression to meaningful areas.

(c) Error distribution: The error map reveals that most regions maintain low absolute errors (dark purple), while slightly higher errors (orange/yellow) appear near coastal boundaries or in sparsely sampled zones. This pattern indicates that the regression component remains smooth and stable within well-sampled regions and only slightly degrades at the edges of the valid domain.

Overall, the three plots collectively demonstrate that the combined model $h(\mathbf{x})$ successfully merges classification and regression into a piecewise-smooth function: it preserves high spatial accuracy in the classification stage and maintains low prediction errors in the regression stage, producing coherent and interpretable outputs across the domain.

Conclusion:

The piecewise regression model $h(\mathbf{x})$ successfully integrates classification and regression components. It behaves smoothly and accurately in valid regions while maintaining distinct boundaries for invalid areas, effectively realizing a piecewise-smooth function across the geographic domain.

Note: The statement has been revised fluently by Chatgpt