Programming assignment Week 6

Question

Download O-A0038-003.xml. and convert then into classification and regression data set. Then use Gaussian Discriminant Analysis (GDA) to build a classification model and build a regression model that represents a piecewise smooth function.

About classification model

(This report used chatgpt for assistants to complete the code)
First, we convert data into classification data set called A. Since the output result is either 0 or 1. We assume $p(x \mid y = k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} exp(-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k))$ for $y \in \{0, 1\}$. By Bayes' rule, $P(y = k \mid x)$ proportional to $P(y = k)p(x \mid y = k)$. We can estimate μ_k , Σ_k and $P(y = k)$ with maximum likelihood estimation. The log-posterior for class k is proportional to $\delta_k(x) = -\frac{1}{2}ln \mid \Sigma_k \mid -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) + lnP(y = k)$. Then we classify:

$$C(\vec{x}) = \begin{cases} 1, & \text{if } \delta_1(\vec{x}) > \delta_0(\vec{x}) \\ 0, & \text{otherwise} \end{cases}$$

About regression model

The classifier first decides whether a data point is valid. If it's valid, use the regressor's prediction. If not valid, output a missing marker value. So we define a function $h(x)$. Use np.where to make sure combine the function correctly.

Result

```
QDA (manual) Accuracy: 0.8252487562189055
Regression MSE: 5.0017
Regression RMSE: 2.2365
R² score: 0.8611
```

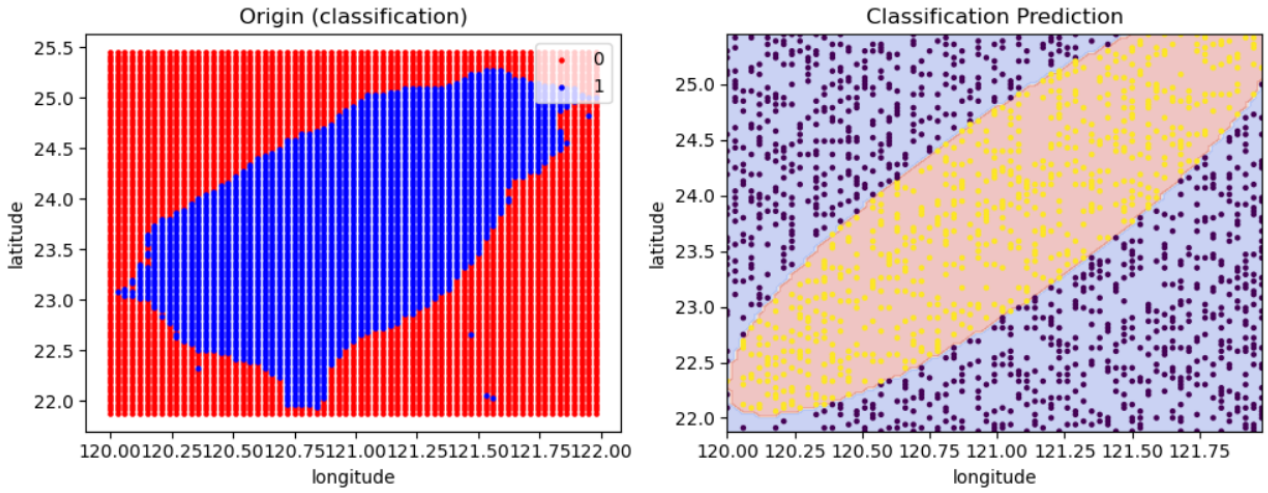

Figure 1: Classification comparison

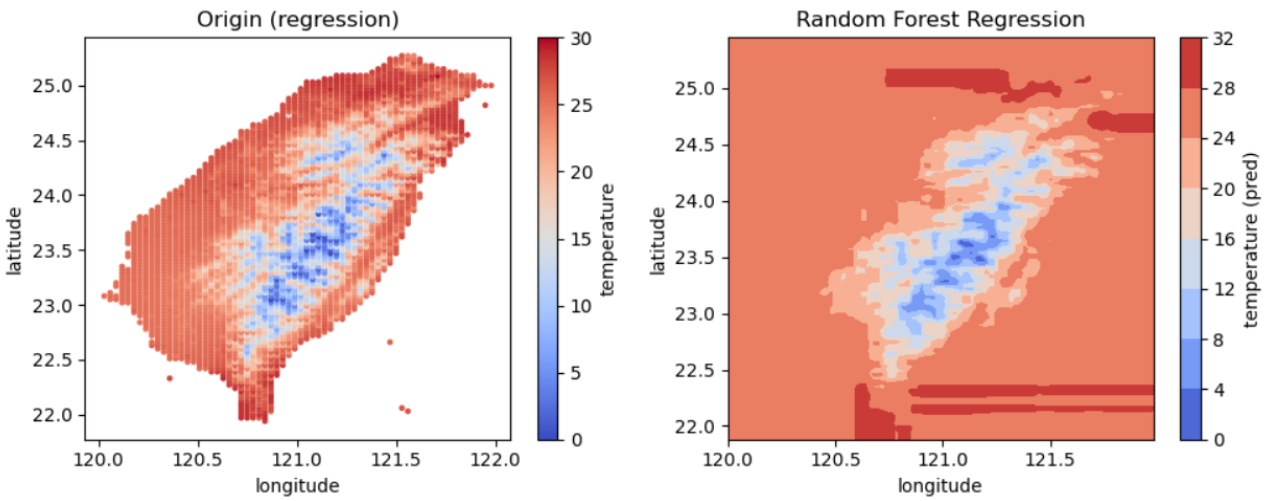

Figure 2: Regression comparison

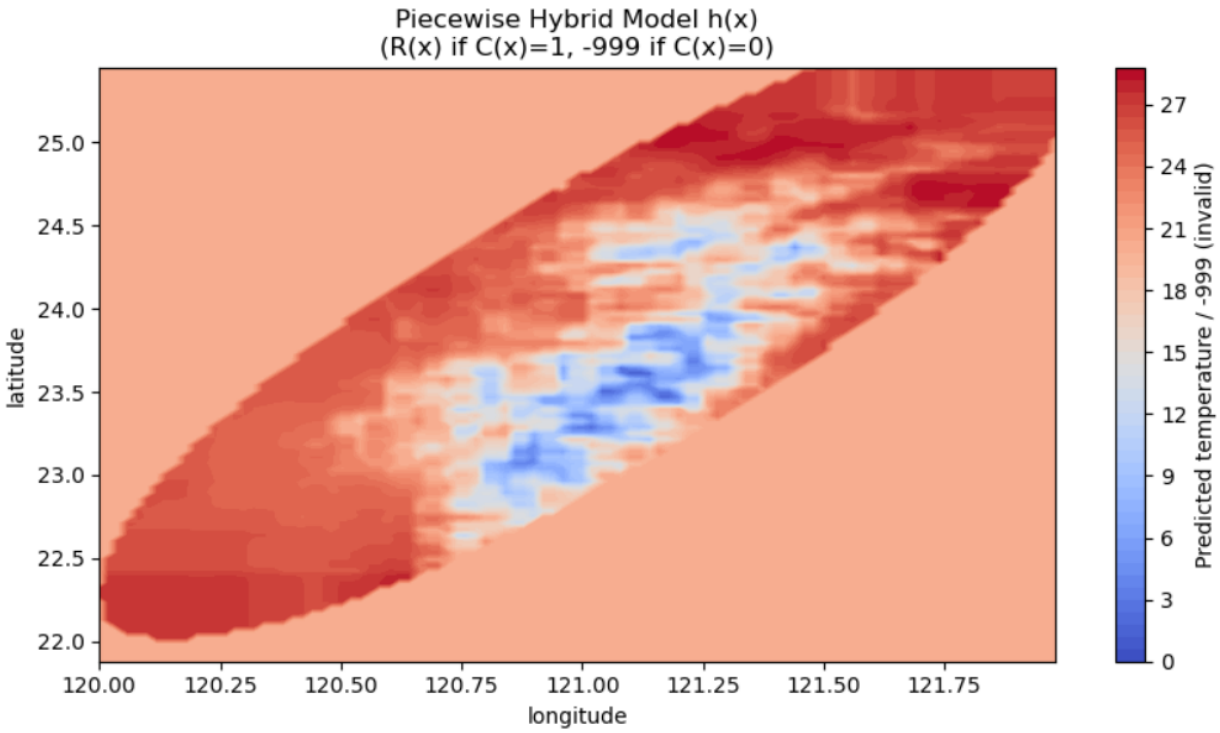

Figure 3: Combine the two models

Analyze result

The QDA accuracy is 0.8252, which is a lot better than using logistic regression. But I think it's not good enough. After all, the shape of Taiwan is still have high different with ellipse. As for the combine of two models, it successfully integrates classification-driven masking with regression prediction, resulting in a physically meaningful and interpretable map.