Machine Learning

313704071_Assignment 6.

1. Classification using GDA

From the assignment 4, we know that since the dataset includes outlying islands and the boundary between the two categories (0 and 1) is irregular—where class 1 is surrounded by class 0—the Taiwan climate dataset is **a linearly inseparable problem**, making LDA unsuitable. Taiwan's main island has a sweet-potato-like shape; therefore, we apply Quadratic Discriminant Analysis (QDA), where the decision boundary is a quadratic surface, and maximum likelihood estimation (MLE) is used for parameter estimation.

Hypothesis: for any class $k \in \{0,1\}$,

$$x \mid y = k \sim N(\mu_k, \Sigma_k)\text{ , where}$$

$$\mu_k\text{: mean, } \Sigma_k\text{: covariance matrix.}$$

QDA allows $\Sigma_0 \neq \Sigma_1$, so the decision boundary is quadratic curve.

By Bayes' Theorem, $P(y = k \mid x) \propto P(x \mid y = k)\pi_k$, where $\pi_k = P(y = k)$.|

Hence, use likelihood function,

$$L(\theta) = P(x \mid y = k)\pi_k = \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)} \pi_k$$

$$\ln L(\theta) = \sum_{i=1}^{n} \left[-\frac{1}{2}\ln|\Sigma_\mathrm{k}| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_\mathrm{k}^{-1}(x_i - \mu_k) + \ln\pi_k \right] + C$$

$$\ln L(\theta)_k = \delta_k(\theta) = \sum_{i=1}^{n} \left[-\frac{1}{2}\ln|\Sigma_\mathrm{k}| - \frac{1}{2}(x_i - \mu_k)^T \Sigma_\mathrm{k}^{-1}(x_i - \mu_k) \right] + n_k \ln\pi_k$$

$$\theta^* = arg \max_{\theta} \delta_k(\theta)$$

We know that by maximum likelihood estimation (MLE), for each class k,

$$\hat{\mu} = \frac{1}{n}\sum x_i,\ \hat{\Sigma} = \frac{1}{n}\sum (x_i - \mu)^T (x_i - \mu)$$

There are three steps in the prediction process:

(1) **Split** the data into training and test sets;

(2) **Fit** the training set to estimate the parameters for each class (μ, Σ, π).

(3) **Predict** – Given a test sample x, determine the optimal parameters(θ^*) and

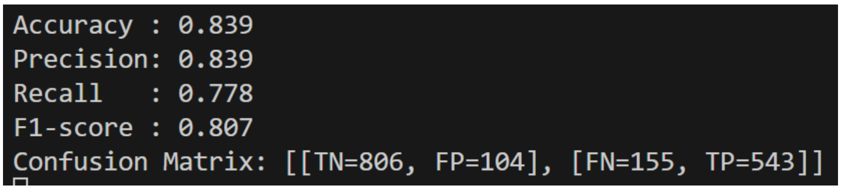
output the predicted value y.

Performance index: Confusion Matrix

TN: prediction 0 when y is 0; **FP**: prediction 1 when y is 0;

FN: prediction 0 when y is 1; **TP**: prediction 1 when y is 1.

$$Accuracy = \frac{TN + TP}{sum\ of\ y_{test}}$$

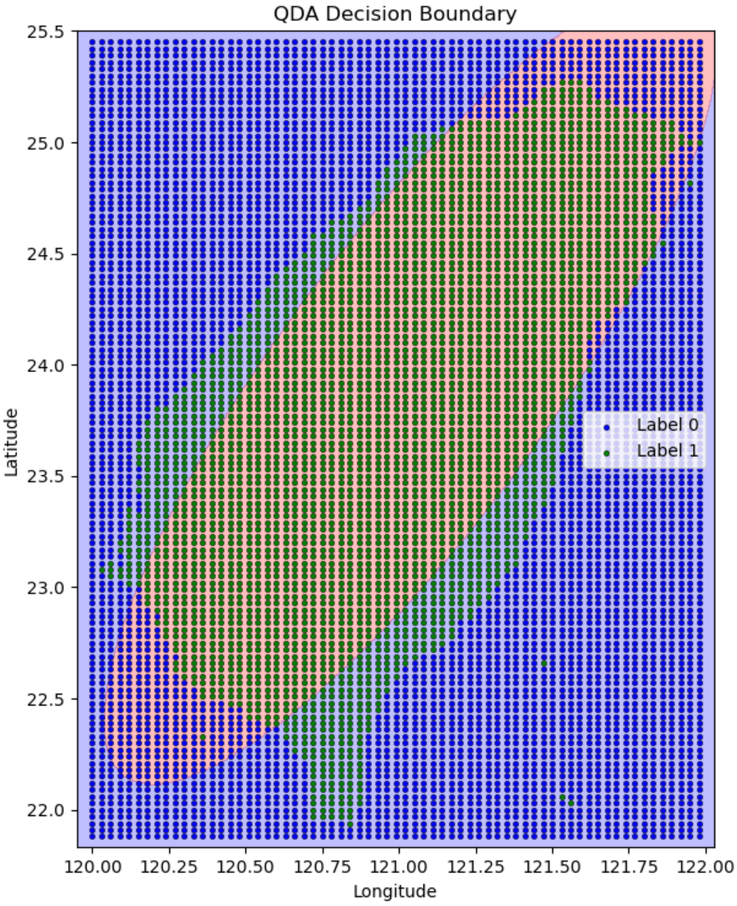


```
Accuracy : 0.839
Precision: 0.839
Recall   : 0.778
F1-score : 0.807
Confusion Matrix: [[TN=806, FP=104], [FN=155, TP=543]]
```

The other three index respectively is precision, recall, and F1-score, where

$$Precision = \frac{TP}{TP+FP},\ Recall = \frac{TP}{TP+FN}, F1\_score = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

Decision boundary: using QDA.



2. Regression – piecewise smooth function

Combine the two models from Assignment 4 into a single function.

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

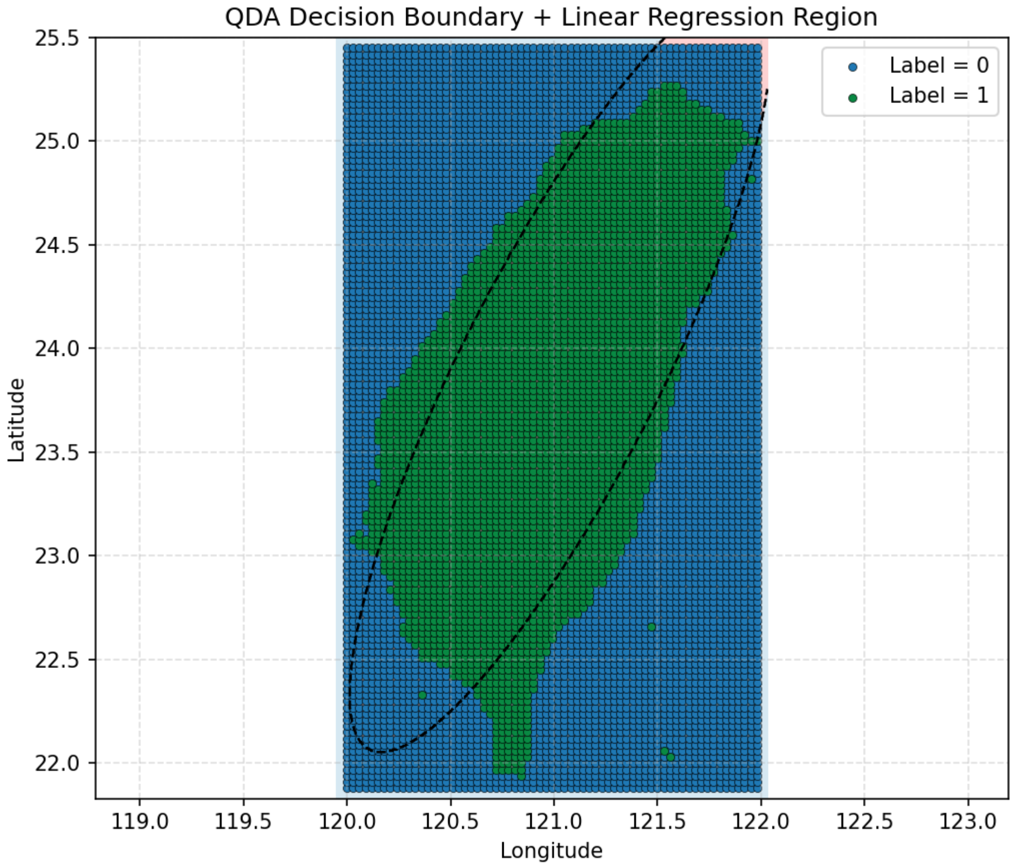
$C(\vec{x})$: classification model, $R(\vec{x})$: regression model.

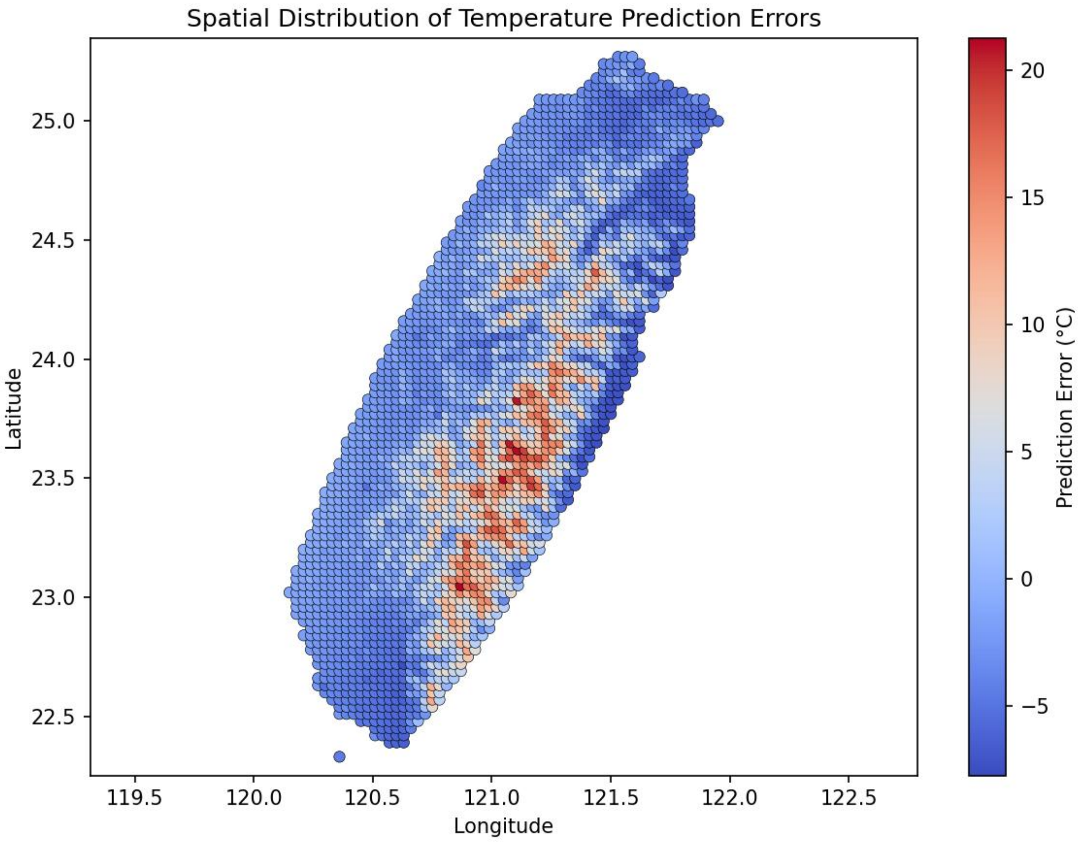
Input: labels.csv, values.csv

***Output**:* Prediction that whether having data using $C(\vec{x})$，if yes (label_pred=1)，continue to predict temperature value using $R(\vec{x})$，Finally, output the predicted values of the temperature of these points.

Brief explanation:

This problem doesn't involve **data splitting**, as the goal isn't generalization evaluation. Instead, we ensure that the classification boundaries are intact, telling the regression model which locations are worth predicting temperature for and which should be skipped. Include plots or tables that demonstrate the behavior of your model. Next, we use linear regression to predict temperature based on latitude and longitude.





Simple linear regression:

$$Temperature = \beta_0 + \beta_1(lon) + \beta_2(lat)$$

This regression assumes that the relationship between temperature and (lon, lat) is roughly linear, where β_1 represents the temperature change for every 0.03 degrees eastward, and β_2 represents the temperature change for every 0.03 degrees northward.

Result:

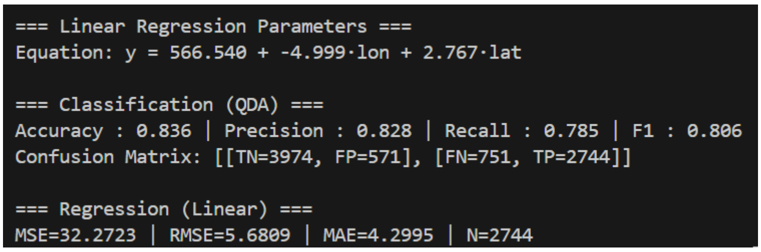
$$Temperature = 566.540 - 4.999(lon) + 2.767(lat)$$

The equation suggests that temperature decreases eastward and increases northward. However, this contradicts the general expectation that temperatures decrease as latitude increases.

Error in prediction: MSE

As shown in **figure** above, red dots represent locations with larger error values. It's not hard to see that the pattern of red dots resembles the distribution of mountain ranges in Taiwan. Because the error is the difference between the predicted value and the actual value, this result is reasonable.

In the future, nonlinear relationships should be considered, for example using polynomial regression.



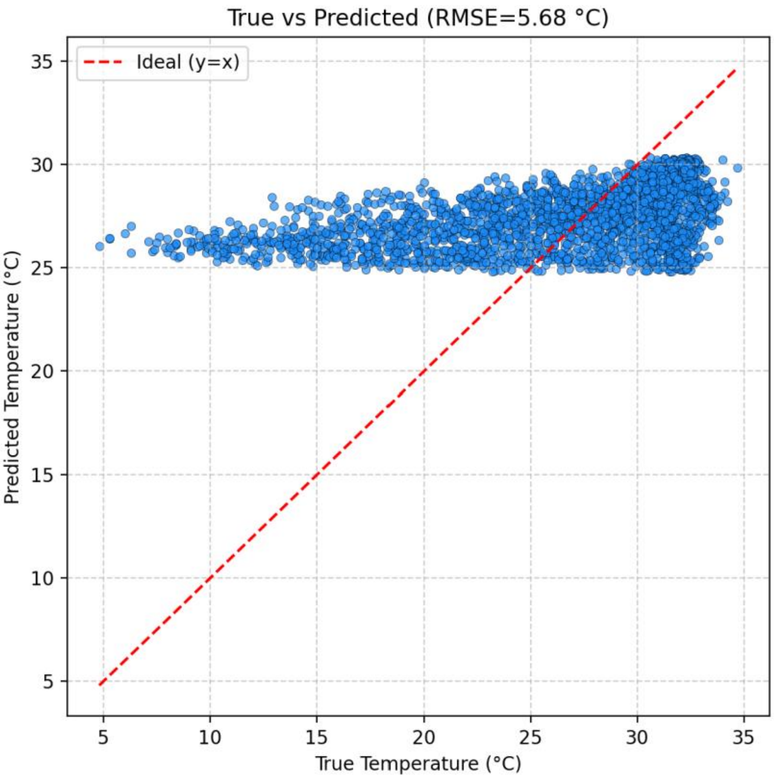
```
=== Linear Regression Parameters ===
Equation: y = 566.540 + -4.999·lon + 2.767·lat

=== Classification (QDA) ===
Accuracy : 0.836 | Precision : 0.828 | Recall : 0.785 | F1 : 0.806
Confusion Matrix: [[TN=3974, FP=571], [FN=751, TP=2744]]

=== Regression (Linear) ===
MSE=32.2723 | RMSE=5.6809 | MAE=4.2995 | N=2744
```

Scatter plot:

Red line: true temp = predicted temp



The performance of prediction model is not good enough.