Implementation of Gaussian Discriminant Analysis

Ou Chia-Yu

October 15, 2025

1 Dataset Description

We use a dataset named `classification_data.csv`, which contains three columns:

- `lon`: longitude
- `lat`: latitude
- `label`: class label (0 or 1)

The dataset is split into 80% training and 20% testing sets.

2 Gaussian Discriminant Analysis (GDA) Theory

Gaussian Discriminant Analysis is a generative model that classifies data based on Bayes' theorem:

$$p(y = k \mid x) \propto p(x \mid y = k) \cdot p(y = k)$$

Where:
- $p(y = k)$: the prior probability for class k - $p(x \mid y = k)$: the likelihood, modeled by a multivariate Gaussian distribution:

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

Here, μ_k is the mean vector and Σ_k is the covariance matrix of class k.

3 Model Implementation and Accuracy

We implemented a custom GDA classifier using Python, as a class named `SelfGDA`. The training process includes:

- Calculating the mean vector and covariance matrix for each class

- Using the multivariate Gaussian density function to compute likelihoods
- Applying Bayes' rule to derive posterior probabilities
- Selecting the class with the highest posterior probability

The accuracy on the test set is:

$$\text{Accuracy} = \boxed{82.40\%}$$

4 Decision Boundary Visualization

Below is a visualization of the decision boundary learned by our GDA model:

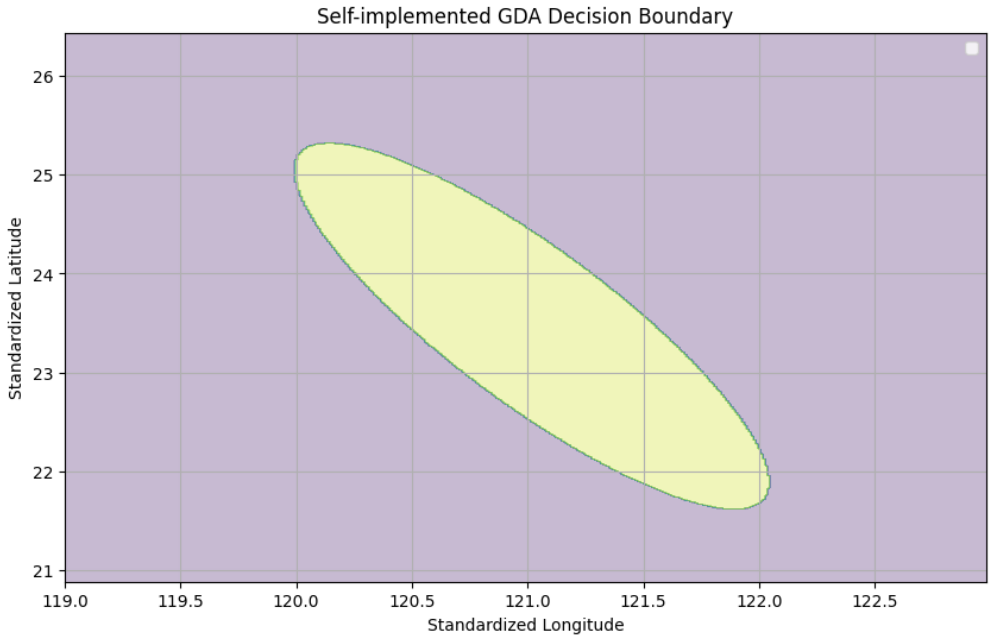

Figure 1: Decision boundary generated by our self-implemented GDA model

The figure plots the classification regions in the longitude-latitude feature space.

5 Conclusion

In this experiment, we implemented a Gaussian Discriminant Analysis (GDA) classifier from scratch and applied it to a binary classification task using geospatial features (longitude and latitude).

Our model achieved a test accuracy of **82.4%**. While this performance is not perfect, it still demonstrates the effectiveness of GDA in capturing the underlying structure of the data using only two features. The decision boundary clearly reflects the separation of the two classes based on spatial information.