ML Programming assignment VI

Name：林欣妤
Student number：314652023

October 13, 2025

1 Description of the Problem

In this assignment, we need to build a classification model using **Gaussian Discriminant Analysis (GDA)**. Specifically, you are required to implement the GDA algorithm from scratch without relying on any built-in classification functions. The GDA model should be trained to distinguish between two classes based on the provided dataset, which contains two-dimensional features (longitude and latitude) and a binary label indicating valid or invalid data points.

2 Gaussian Discriminant Analysis (GDA) for Classification

2.1 Model Implementation

We implemented **Gaussian Discriminant Analysis (GDA)** each class has its own covariance matrix. The model estimates:

- Prior probabilities: $P(y = 1) = \phi$, $P(y = 0) = 1 - \phi$.
- Class means: μ_0, μ_1.
- Class covariance matrices: Σ_0, Σ_1.

Given a sample $\mathbf{x}$, the posterior probabilities are computed using Bayes' theorem:

$$\ln P(y = k|\mathbf{x}) \propto \ln P(\mathbf{x}|y = k) + \ln P(y = k),$$

where $P(\mathbf{x}|y = k)$ is modeled as a multivariate Gaussian distribution:

$$\mathbf{x}|y = k \sim \mathcal{N}(\mu_k, \Sigma_k), \quad k = 0, 1.$$

The predicted class is determined by selecting the class with the higher posterior probability.

2.2 Training and Performance

The GDA model was trained on the training set. On the test set, predictions were generated and the accuracy was computed:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of samples}}$$

The test accuracy achieved by the model is:

$$\text{Accuracy} = 0.9900$$

2.3 Decision Boundary

To visualize the model behavior, we created a dense grid of points covering the feature space. Each point was classified using the trained GDA model.

- The decision regions were plotted with semi-transparent colors: blue for class 0 and red for class 1.
- The decision boundary was highlighted with a black dashed line, representing the locus where:

$$\ln P(\mathbf{x}|y=1) + \ln P(y=1) = \ln P(\mathbf{x}|y=0) + \ln P(y=0).$$

- Test data points were overlaid as black-edged crosses to show their true labels.

The results are shown in Figure 1.

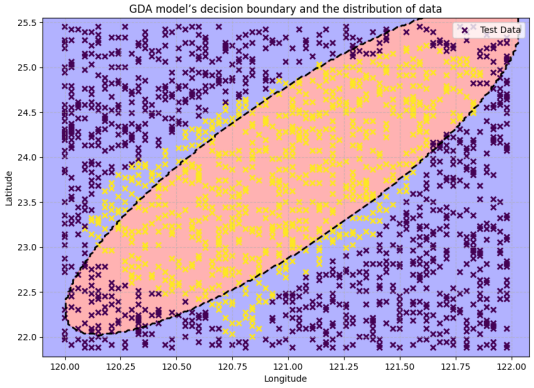

Figure 1: Decision Boundary of GDA model

References

[1] Chat-GPT (Apply GPT to revise and correct the English content of the report, and ask about some programming techniques)

[2] Google Gemini (Apply Gemini to revise and correct the English content of the report, and ask about some programming techniques)