1141ML-Week 6 Programming Assignment 01

Taiwan Weather Dataset Prediction

313652018 王宣瑋

September 2025

Tasks

Using Gaussian Discriminant Analysis (GDA) to build a classification model for the assignment 4 dataset.

a) code

code file name: 1141ML-Week-6-Prog01-313652018

b) Explnantion of GDA

Gaussian Discriminant Analysis (GDA) is a **generative classification model** that assumes the feature vector $x \in \mathbb{R}^d$ is drawn from a class-conditional Gaussian (normal) distribution. For a binary classification task where $y \in \{0, 1\}$, the model assumes

$$p(x \mid y = k) = \mathcal{N}(x; \mu_k, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_k)^\top \Sigma^{-1}(x - \mu_k)\right),$$

where

- μ_k is the mean vector of class k,
- Σ is the shared covariance matrix across classes.

The prior probability of class $y = 1$ is denoted by

$$\phi = P(y = 1), \quad P(y = 0) = 1 - \phi.$$

By Bayes' theorem, the posterior probability that a sample x belongs to class 1 is given by

$$P(y = 1 \mid x) = \frac{p(x \mid y = 1)\, P(y = 1)}{p(x \mid y = 1)P(y = 1) + p(x \mid y = 0)P(y = 0)}.$$

The GDA classifier predicts the label as

$$\hat{y} = \begin{cases} 1, & \text{if } \log \frac{P(y=1|x)}{P(y=0|x)} > 0, \\ 0, & \text{otherwise.} \end{cases}$$

When both classes share the same covariance matrix Σ, the decision boundary $\log P(y = 1 \mid x) - \log P(y = 0 \mid x) = 0$ is **linear** in x. If different covariances Σ_0 and Σ_1 are used, the boundary becomes **quadratic**.

—

In this task, each sample $x = (\text{latitude}, \text{longitude})^\top$ represents a spatial location, and the target $y \in \{0, 1\}$. Hence the GDA model works. This is a intuitive statistic way to solve the classfication problem.

c) Accuracy

The Gaussian Discriminant Analysis (GDA) model was trained on the given dataset, where each sample $x = (\text{latitude}, \text{longitude})^\top$ represents a spatial location, and the target label $y \in \{0, 1\}$ indicates whether the temperature measurement is valid (1) or invalid (0).

To evaluate model performance, the dataset was divided into 75% for training and 25% for testing. In addition, a 25-fold cross-validation procedure was performed to obtain a more robust estimate of generalization accuracy.

The classification accuracy is defined as

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

The test set accuracy of the GDA model is 53.4%, and the 25-fold cross-validation mean accuracy is 53.38%, indicating that the model performs only slightly better than random guessing.

The confusion matrix on the test set is shown as the following table.

	Predicted 0	Predicted 1
True 0	1074	6
True 1	930	0

Table 1: Confusion matrix for the test set.

This result suggests that the model tends to classify nearly all samples as class 0, leading to poor recall for the valid (class 1) data points. Such behavior is likely caused by class imbalance and the linear boundary assumption of GDA.

d) Plot decision boundary

To visualize the classification behavior, the decision boundary of the trained GDA model was plotted in the two-dimensional feature space. Longitude is plotted on the x-axis and latitude on the y-axis, so that the map orientation corresponds to real-world geography.

Each point represents a location in the dataset: orange points correspond to valid samples ($y = 1$), and blue points correspond to invalid samples ($y = 0$). The shaded background shows the model' s predicted regions for each class.

Figure 1 below illustrates the learned decision boundary of the GDA model.

The boundary is nearly linear, reflecting the shared covariance assumption. However, it fails to capture the complex geographic shape of the valid region.

This visualization confirms that the GDA' s linear discriminant function

$$g(x) = w^\top x + b = 0,$$

is insufficient to separate the two spatial clusters effectively. Future improvements may include using Quadratic Discriminant Analysis (QDA),non-linear kernel models, or incorporating additional spatial featuressuch as altitude or regional temperature patterns.

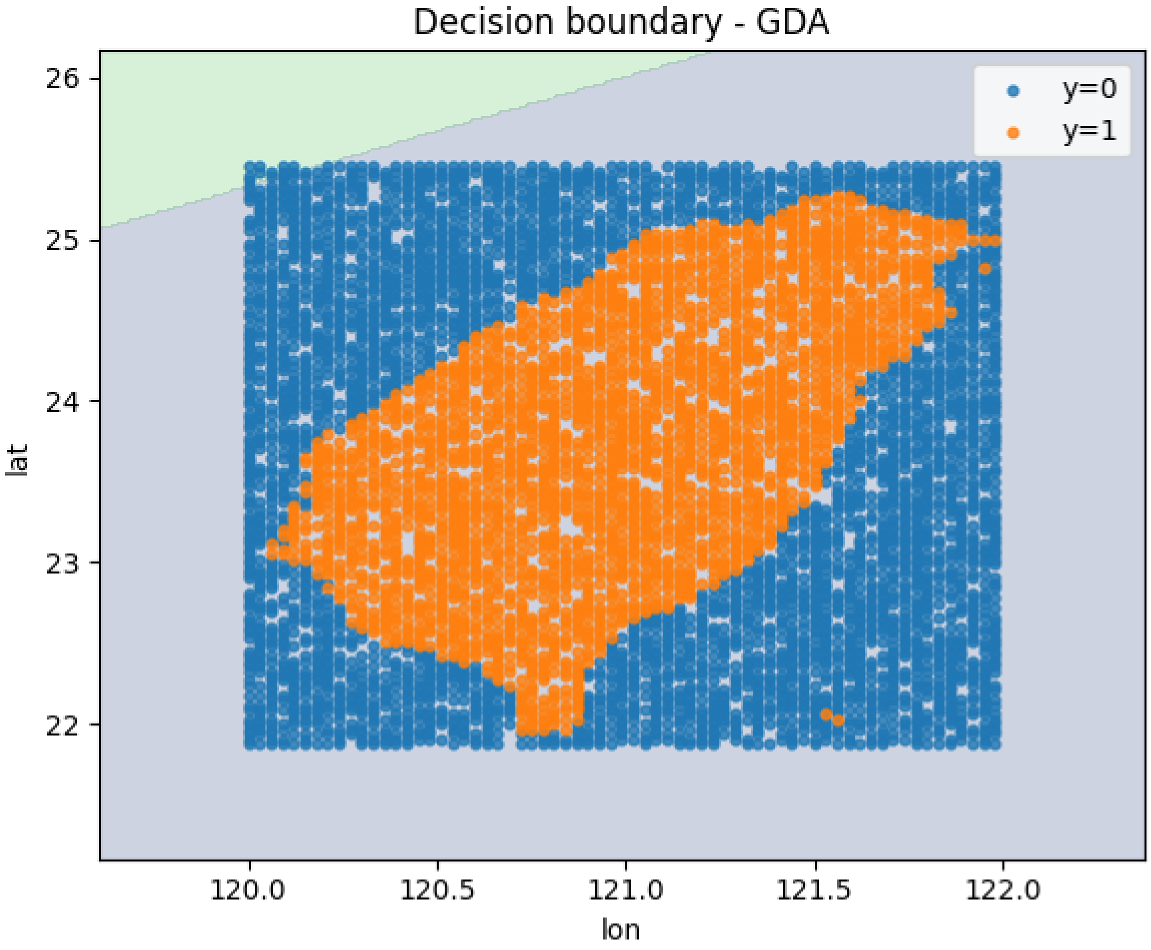

Figure 1: Decision boundary produced by the GDA classifier.