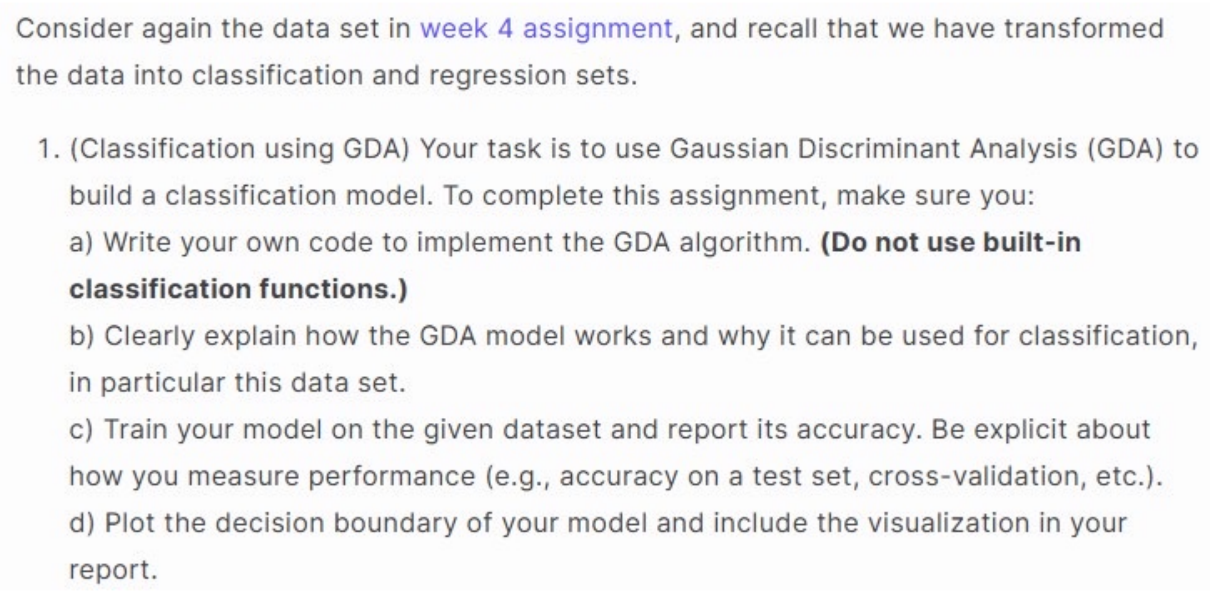


Consider again the data set in week 4 assignment, and recall that we have transformed the data into classification and regression sets.

1. (Classification using GDA) Your task is to use Gaussian Discriminant Analysis (GDA) to build a classification model. To complete this assignment, make sure you:
   a) Write your own code to implement the GDA algorithm. **(Do not use built-in classification functions.)**
   b) Clearly explain how the GDA model works and why it can be used for classification, in particular this data set.
   c) Train your model on the given dataset and report its accuracy. Be explicit about how you measure performance (e.g., accuracy on a test set, cross-validation, etc.).
   d) Plot the decision boundary of your model and include the visualization in your report.

a) code 在另外的文字檔

b) 假設在每個類別 $\overset{\in \mathbb{R}}{y}$ 的資料 $\overset{\in \mathbb{R}^n}{x}$ 都是從 Gaussian distribution 中得到的 (此以分成2類說明)

ie,

第k個類別

$$P(x|y=k) = N(\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left(-(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right), \quad k=0,1$$

且 $P(y) = \phi^y (1-\phi)^{1-y}$

(若 $\Sigma_0 = \Sigma_1$, 則是 LDA

若 $\mu_0 \neq \mu_1$ or $\Sigma_0 \neq \Sigma_1$, 則是 QDA)

將一部分資料 (training data) : $\{x^{(i)}, y^{(i)}\}$ 代入 likelihood function $L(\theta) = \prod_{i=1}^{m} P(x^{(i)}, y^{(i)})$

接著求出 $\theta^* = \underset{\theta}{\operatorname{argmax}}\, L(\theta) = \underset{\theta}{\operatorname{argmin}} -L(\theta)$, 即為 $\mu_k^*, \Sigma_k^*, \phi^*$, $k=0,1$

則給新的資料 $\tilde{x}$, 則可計算出

$$P(\tilde{x}|y=k) = N(\mu_k^*, \Sigma_k^*)$$
$$P(y) = (\phi^*)^y (1-\phi^*)^{1-y}$$

再透過貝氏定理, 計算

$$P(y=k|\tilde{x}) = \frac{P(\tilde{x}|y=k)P(y=k)}{P(\tilde{x}|y=0)\cdot P(y=0) + P(\tilde{x}|y=1)\cdot P(y=1)}, \quad k=0,1$$

比較 $P(y=0|\tilde{x})$ 和 $P(y=1|\tilde{x})$ 的值.

若 $P(y=0|\tilde{x}) > P(y=1|\tilde{x})$, 則 $\tilde{x}$ 判斷屬於類別 0

若 $P(y=0|\tilde{x}) < P(y=1|\tilde{x})$, 則 $\tilde{x}$ 判斷屬於類別 1

c) 將 HW4 已分成 label 0, label 1 的 8040筆資料分成

training data：5628筆 (70%)， test data：2412筆 (30%)

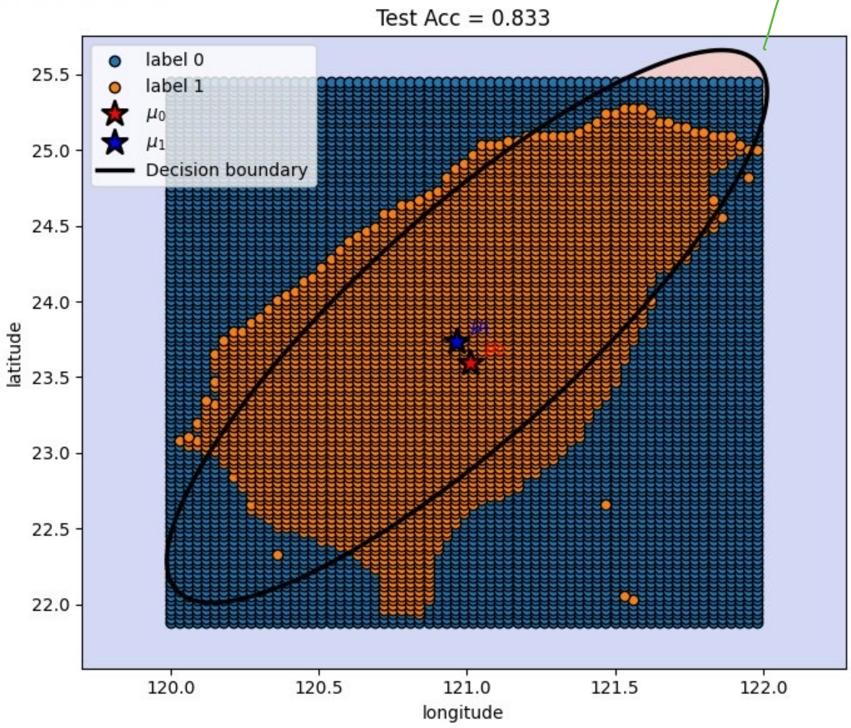
此沒有 stop criteria，所以不須要 validation data

此 model 的 accuracy 用 test accuracy $= \dfrac{\text{正確預測的 test data 數量}}{\text{test data 數量}}$ 表示 $= 2412$

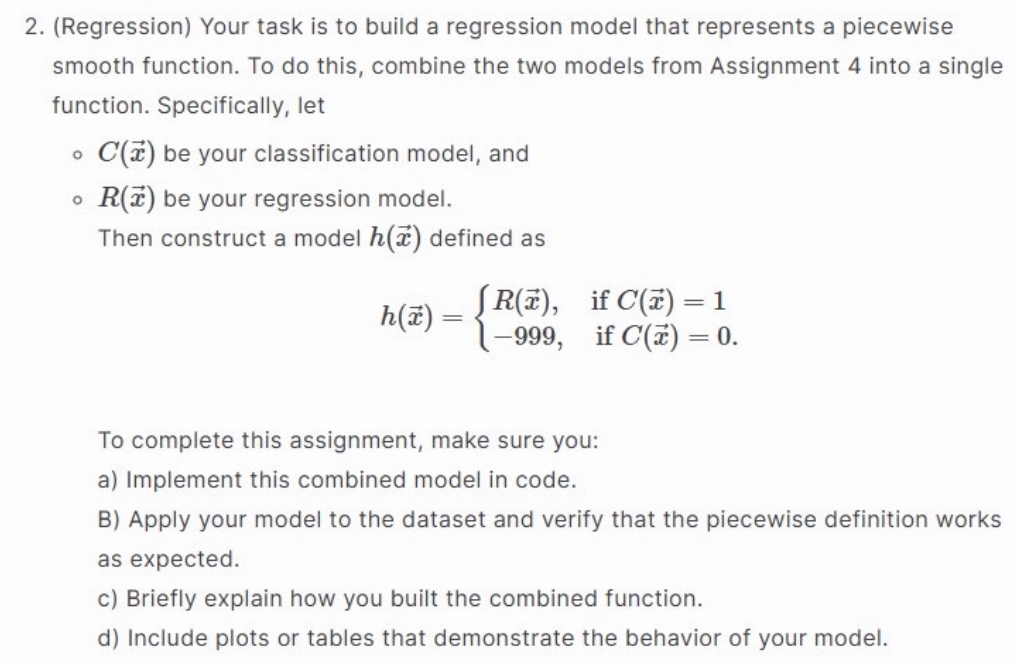
結果為 0.8333 = 83.33%

d)

```
Test Accuracy = 0.8333
```



→ Decision boundary



2. (Regression) Your task is to build a regression model that represents a piecewise smooth function. To do this, combine the two models from Assignment 4 into a single function. Specifically, let

   - $C(\vec{x})$ be your classification model, and
   - $R(\vec{x})$ be your regression model.

   Then construct a model $h(\vec{x})$ defined as

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

To complete this assignment, make sure you:

a) Implement this combined model in code.

B) Apply your model to the dataset and verify that the piecewise definition works as expected.

c) Briefly explain how you built the combined function.

d) Include plots or tables that demonstrate the behavior of your model.

a) code 在另外的文字檔

b) c) d)

在 classfication_dataset 中的 8040筆資料，取 80% 的 data (6432筆) 當作 training data，接著利用 GDA 訓練出 classification model $C(\vec{x})$，將剩下的 20% 的 test data (1608筆) 用來測試 $C(x)$，得到準確率為 83%，

再從 regression_dataset 中的 3495筆 data，取 80% 的 data (2796筆) 當作 training data，接著利用線性迴歸訓練出 regression model $R(\vec{x})$，剩下的 20% 的資料 (699筆) 則是用來測試 model $h(\vec{x})$

將這 699筆資料代入 model C，結果為 521筆被判斷為 label 1，178筆被判斷為 label 0，以被判斷 label 1 的 521筆 data 來測試 $R(\vec{x})$ 的表現。

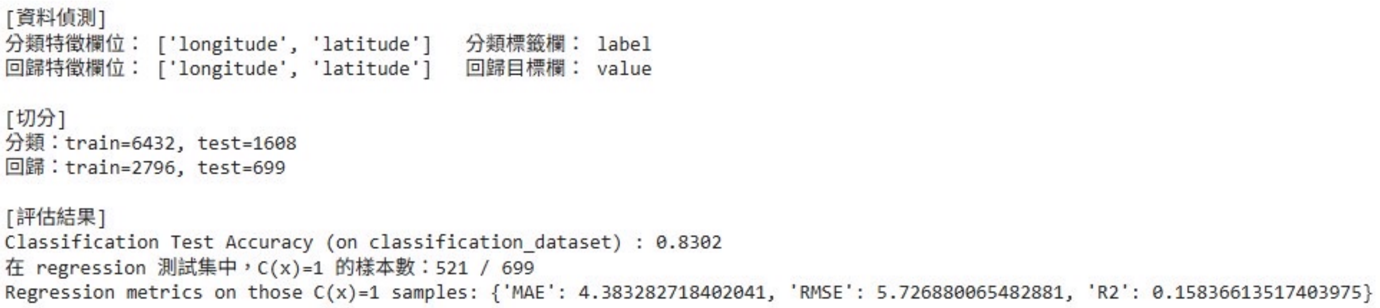
可得到 Mean Absolute Error (MAE) $= \frac{1}{521}\sum_{i=1}^{521}|y_i-\hat{y}_i| = 4.38$

Root Mean Squared Error (RMSE) $= \sqrt{\frac{1}{521}\sum_{i=1}^{521}(y_i-\hat{y}_i)^2} = 5.73$

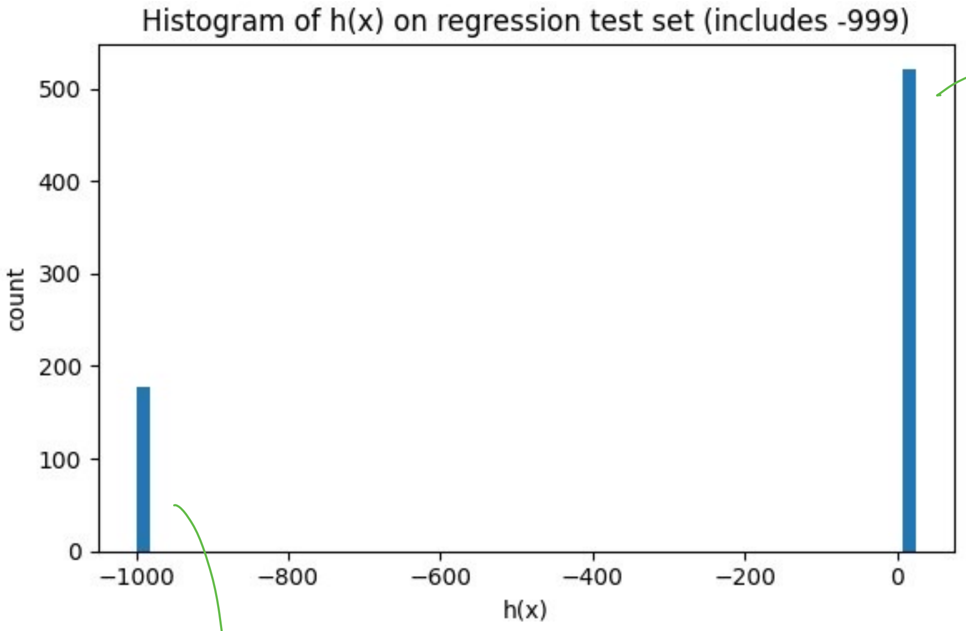
Coefficient of Determination $R^2 = 1 - \frac{\sum_{i=1}^{521}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{521}(y_i-\bar{y})^2} = 0.158$, where $\bar{y}$ 是真實值的平均

越接近 1 越好

代表 $R(\vec{x})$ 可以解釋大約 15.8% 的 y 的變化



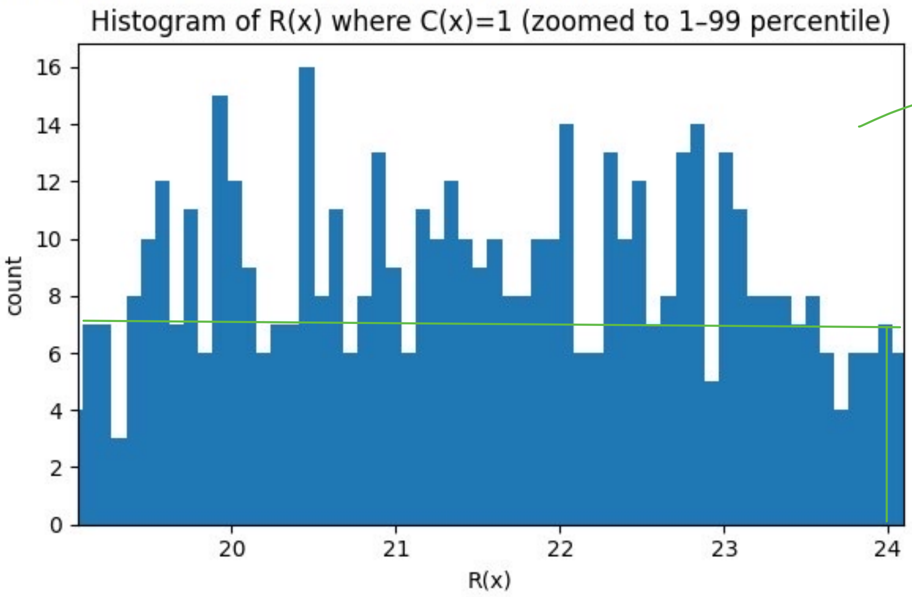
```
[資料偵測]
分類特徵欄位： ['longitude', 'latitude']   分類標籤欄： label
回歸特徵欄位： ['longitude', 'latitude']   回歸目標欄： value

[切分]
分類：train=6432, test=1608
回歸：train=2796, test=699

[評估結果]
Classification Test Accuracy (on classification_dataset) : 0.8302
在 regression 測試集中，C(x)=1 的樣本數：521 / 699
Regression metrics on those C(x)=1 samples: {'MAE': 4.383282718402041, 'RMSE': 5.726880065482881, 'R2': 0.15836613517403975}
```



699筆data中被model C
判斷為1的data有521筆

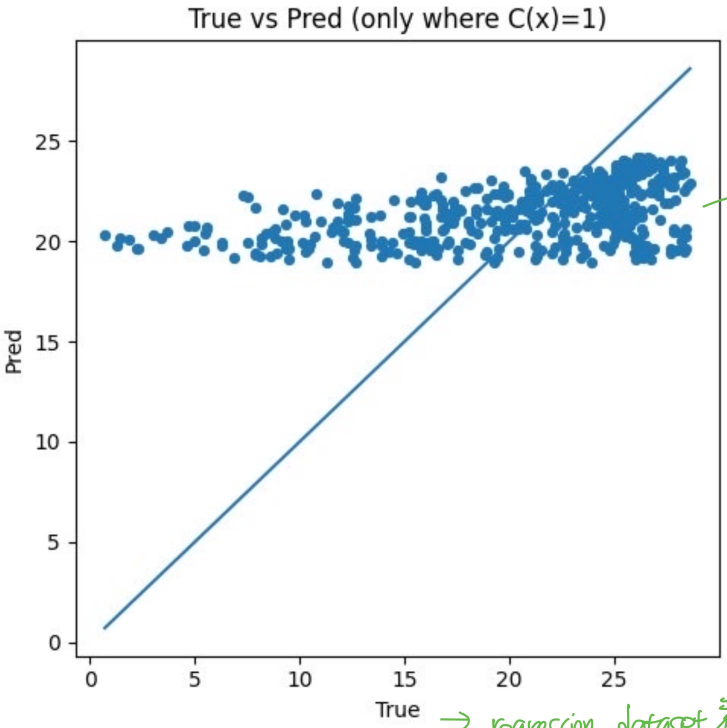
699筆data中被model C
判斷為0的data有178筆

```
[連續分佈範圍參考] 1% ~ 99% 百分位: [19.063, 24.089]
```



將上圖右邊長條形放大來看.

在521筆data中, 有7筆data
的h(x)值為24



model R($\vec{x}$)的值 ←
(預測)

→ 藍點為被判斷 C($\vec{x}$)=1 的521筆 data

可看出 model h 預測所有的data 的值都差不多 (19~24之間)

表示 model 只學到一個接近平均的結果

→ regression_dataset 裡真實的值

```
[Piecewise 行為檢查]
h(x) 中等於 -999 的個數：178
h(x) 中不等於 -999 的個數：521 (這些點是由 R(x) 給的)
```