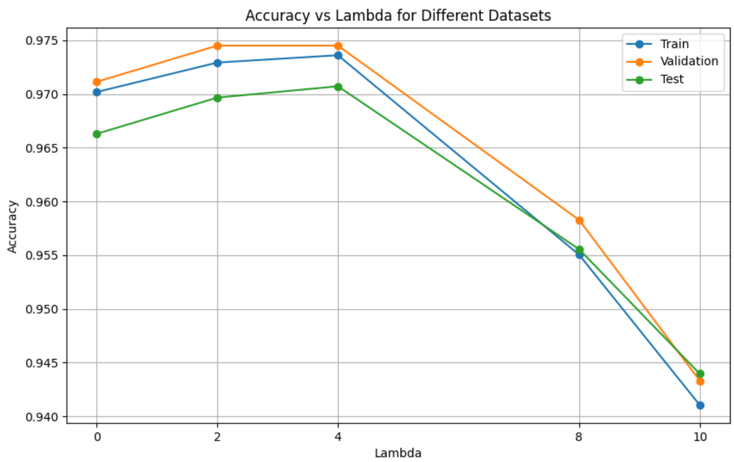
# IML Project 3 Report - Nir Ellor

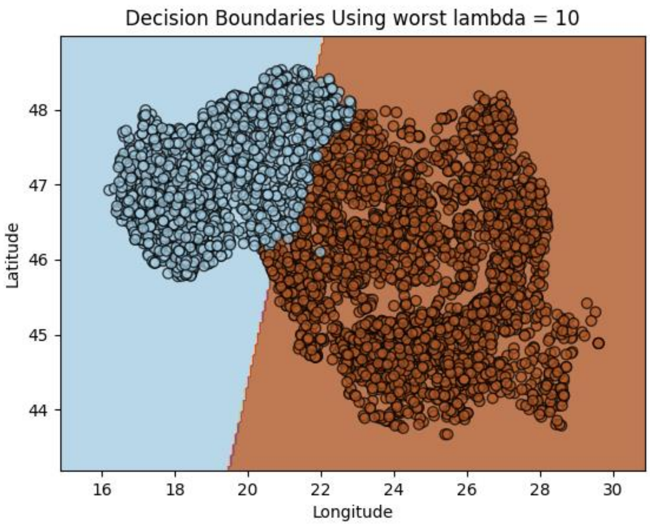
**3.2.1:**

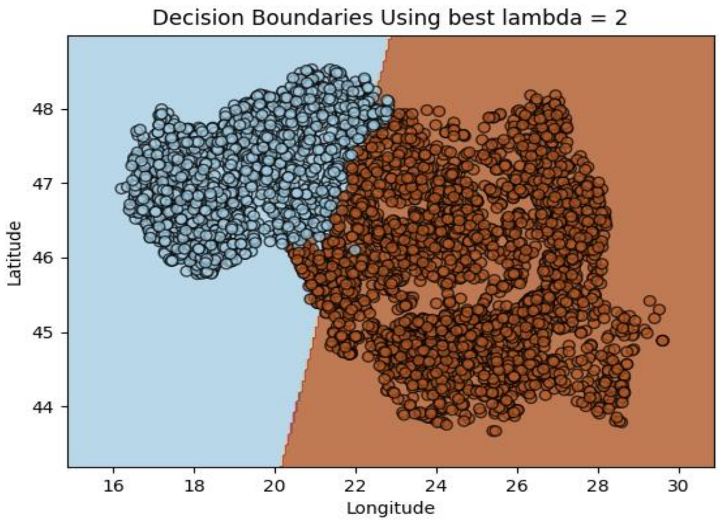
| λ | Validation Accuracy | Test Accuracy |
|---|---|---|
| 0 | 0.971134 | 0.966295 |
| 2 | 0.974505 | 0.969665 |
| 4 | 0.974505 | 0.970718 |
| 8 | 0.958281 | 0.955551 |
| 10 | 0.943321 | 0.943965 |




Hence, the test accuracy of the best model according to the validation set is 0.969665, and best λ is 2.

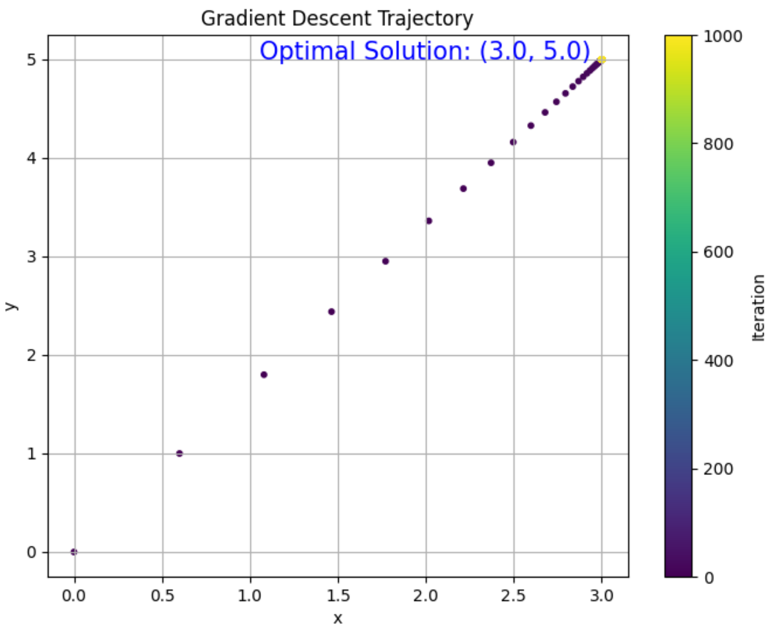
**3.2.2:**



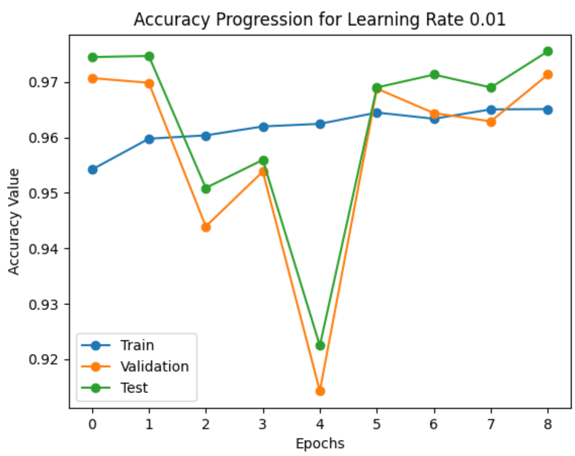

We can observe that the λ parameter affect the algorithm especially in the dots between (20,46) and (20,48), the area where the line classifier is. The difference between the best λ and the worst takes place as the number of blue dots that are in the red area is higher at λ = 10 than λ = 2.

**4.1:**



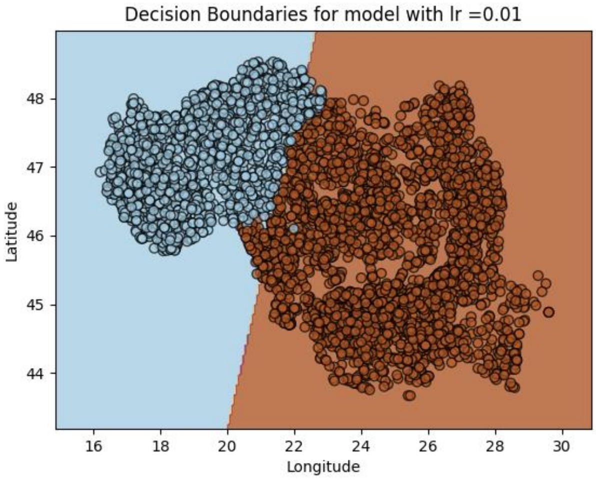

We can observe that the gradient descent implementation stopped at the point (3,5), reaching it as more iterations passed. This makes sense since f gets its minimal value at that point.

**6.3.1:**

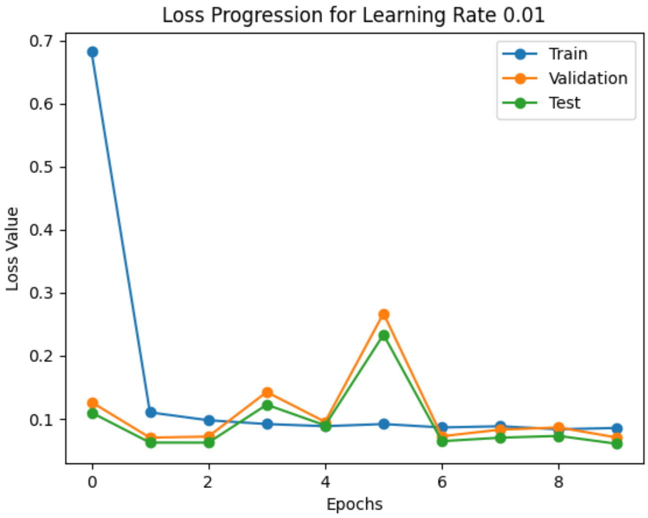



| Learning rate | Validation Accuracy | Test Accuracy |
|---|---|---|
| **0.1** | 0.968401 | 0.968401 |
| **0.01** | 0.971350 | 0.975558 |
| **0.001** | 0.970508 | 0.975980 |

Hence, the test accuracy of the best model according to the validation set is 0.975558, and the best learning rate is 0.01.




**6.3.2:**




The model generalized well from the data. The validation and test accuracies for all learning rates are close to each other, indicating minimal overfitting. Specifically, with the best learning rate of **0.01**, the validation accuracy was **97.13%**, and the corresponding test accuracy was **97.56%**, showing strong generalization to unseen data. The stability of the loss and accuracy curves further supports this conclusion. This even though the first epoch of training loss shows a significant difference, the model still generalized well. Early differences, especially in the first epoch, might occur because the model starts with random weights and adjusts as training progresses.

## 6.3.3:

The plot using λ = 2 (ridge regression) produces a smoother and more generalized decision boundary, avoiding overfitting to the training data. In contrast, the plot with learning rate = 0.00 (logistic regression) creates a sharp and more rigid boundary, which may indicate overfitting.   Moreover:

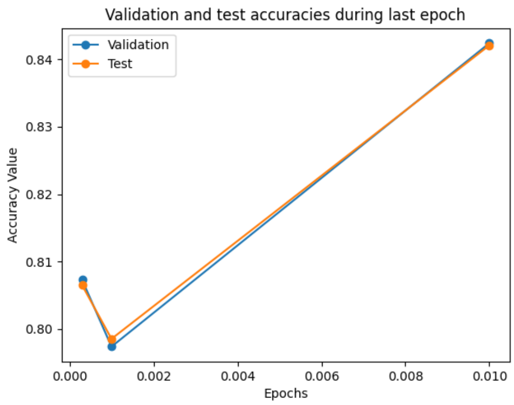
Ridge Regression (λ = 2): Validation accuracy is 97.4% and Test accuracy is 96.9%.

Logistic Regression (learning rate = 0.001): Validation accuracy is 96.9% and Test accuracy is 97.0%.

As a result, both models perform similarly, with only slight differences in validation and test accuracies.

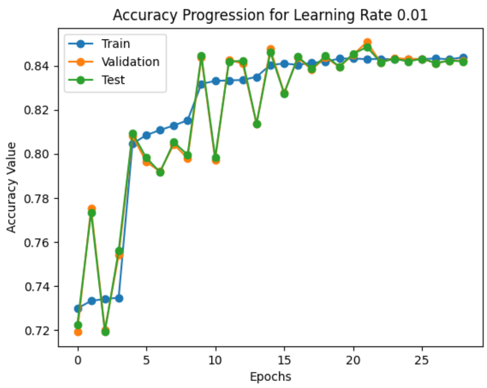
## 6.4.1:

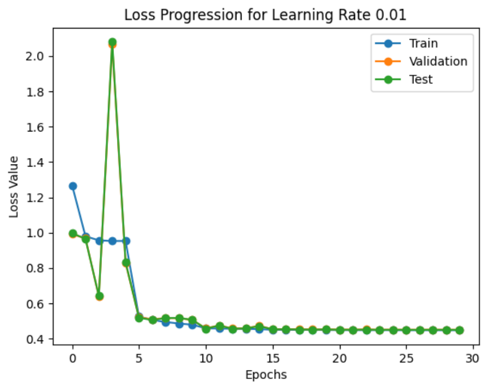
| Learning rate | Validation Accuracy | Test Accuracy |
|---|---|---|
| **0.01** | 0.842398 | 0.842083 |
| **0.001** | 0.797391 | 0.798543 |
| **0.0003** | 0.807293 | 0.806507 |

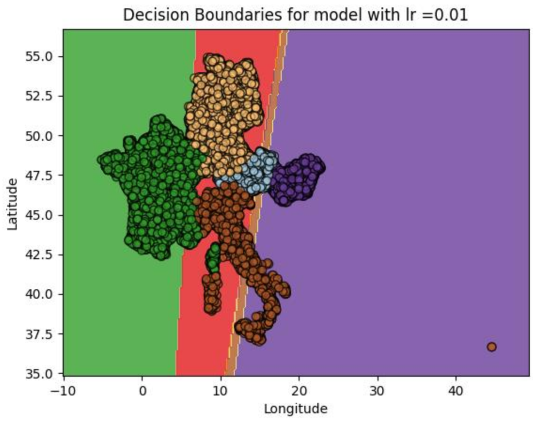



Hence, the test accuracy of the best model according to the validation set is 0.842398, and the best learning rate is 0.01.

## 6.4.2:

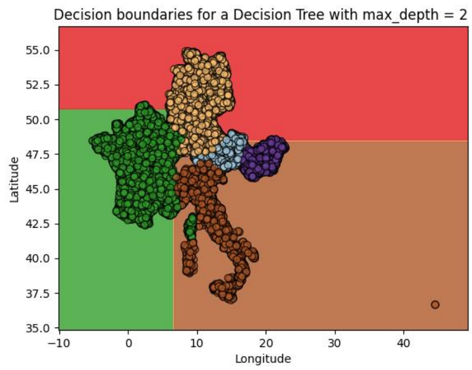




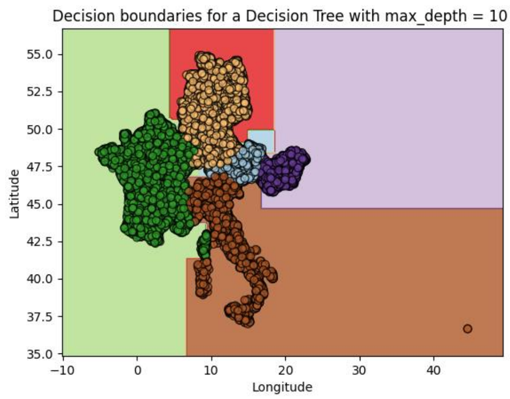

The model generalized well as training, validation, and test accuracies converge closely after initial fluctuations, showing consistent performance. The loss progression aligns across datasets and stabilizes, indicating effective learning without overfitting. The decision boundaries are reasonable and effectively separate the classes. Despite early instability, the model demonstrates strong generalization to unseen data.

**6.4.3:**




The decision tree with depth 2 has an accuracy of 75%. This model is less efficient than the previous logistic regression model mainly because it tries to separate 5 countries using only 4 vertical lines, resulting in lower accuracy.

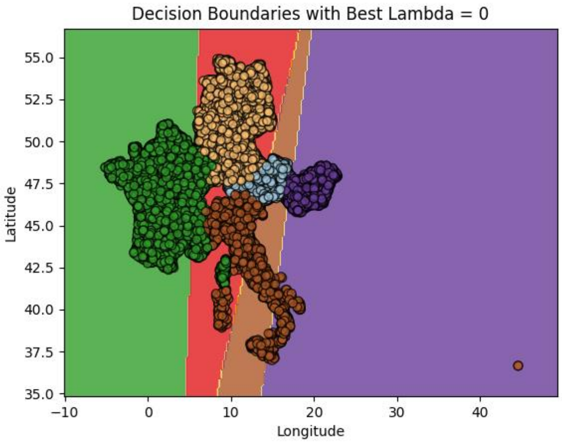
**6.4.4:**




Our answer has changed. The decision tree with depth 10 has an accuracy of 99%. This model is more efficient than the previous logistic regression model, probably due its flexibility and ability to draw many horizontal and vertical lines as its depth are 10. The high number of lines can surpass any diagonal lines bordering the countries, thus achieving almost 100% accuracy.

**6.4.5:**

| λ | Train Accuracy | Validation Accuracy | Test Accuracy |
|---|---|---|---|
| **0** | 0.746326 | 0.770722 | 0.769098 |
| **2** | 0.403114 | 0.419941 | 0.420727 |
| **4** | 0.356409 | 0.347689 | 0.347689 |
| **8** | 0.320793 | 0.369276 | 0.369276 |
| **10** | 0.314886 | 0.347689 | 0.077491 |




Hence, the the best model according to the validation set is with λ = 0.

This model produces 77% accuracy, lower than the multi logistic regression model (84%). In addition, both models generalize well among validation and test, with a slight advantage to the previous model.

**6.4.6:**

France in green, Germany in orange, Austria in blue, Hungary in purple and Italy in Red.

# Theoretical Part – Convexity

## מטלה 3

**.1**

הוכחה: בהינתן $X,Y\in\mathbb{R}^n$ ו- $\theta\in[0,1]$. נשים לב כי מתקיים:

$$f(\theta X+(1-\theta)Y)=a(\theta X+(1-\theta)Y)+b=a\theta X+a(1-\theta)Y+b$$

$$\theta f(X)+(1-\theta)f(Y)=\theta(aX+b)+(1-\theta)(aY+b)=a\theta X+\cancel{b\theta}+aY+b-a\theta Y-\cancel{b\theta}=a\theta X+a(1-\theta)Y+b$$

לכן מהגדרה $f$ קמורה, כנדרש.

**.2**

הוכחה: יהיו $a,b,c\in\mathbb{R}$. נשים לב כי $f$ גזירה בכל $\mathbb{R}^n$ ומתקיים:

$$f'(X)=2aX+b\ ,\ f''(X)=2a$$

מתקיים ש- $a\geq 0 \longleftrightarrow f''(X)\geq 0 \underset{\text{אינפי 2}}{\longleftrightarrow} f$ קמורה ב- $\mathbb{R}^n$. לכן $f$ קמורה ב- $\mathbb{R}^n$ אם"ם $a\geq 0$, כנדרש.

**.3**

הוכחה: נשים לב כי $f$ גזירה פעמיים בכל $\mathbb{R}^n$ ומתקיים $f''(X)=e^X\geq 0$, לכן $f$ קמורה ב- $\mathbb{R}^n$, כנדרש.

**.4**

הוכחה: בהינתן $c,X,Y\in\mathbb{R}^n$ ו- $\theta\in[0,1]$. נשים לב כי מתקיים:

$$f(\theta X+(1-\theta)Y)=\max(\theta X+(1-\theta)Y,c)$$

$$\theta f(X)+(1-\theta)f(Y)=\theta\max(X,c)+(1-\theta)\max(Y,c)=\begin{cases}\theta X+(1-\theta)Y & X\geq c,\ Y\geq c \quad (1)\\ \theta c+(1-\theta)Y & X<c,\ Y\geq c \quad (2)\\ \theta X+(1-\theta)c & X\geq c,\ Y<c \quad (3)\\ \theta c+(1-\theta)c=c & X<c,\ Y<c \quad (4)\end{cases}$$

לכן, - בהינתן (1),

$$f(\theta X+(1-\theta)Y)=\max(\theta X+(1-\theta)Y,c)=\theta X+(1-\theta)Y=\theta f(X)+(1-\theta)f(Y)$$

אחרת,

$$c>\theta X+(1-\theta)Y\geq\theta c+(1-\theta)c=c$$

- בהינתן (4),

$$f(\theta X+(1-\theta)Y)=\max(\theta X+(1-\theta)Y,c)=c=\theta f(X)+(1-\theta)f(Y)$$

- בהינתן (3), אם $\max(\theta X+(1-\theta)Y,c)=\theta X+(1-\theta)Y$, אזי,

$$f(\theta X+(1-\theta)Y)=\theta X+(1-\theta)Y\underset{Y<c}{<}\theta X+(1-\theta)c=\theta f(X)+(1-\theta)f(Y)$$

אחרת, $\max(\theta X+(1-\theta)Y,c)=c$. לכן:

$$f(\theta x+(1-\theta)y) \le \theta f(x)+(1-\theta)f(y) \longleftrightarrow c \le \theta x+(1-\theta)c \longleftrightarrow \theta c \le \theta x$$

(אם $\theta=0$ אז: $f(\theta x+(1-\theta)y) = f(y) = \theta f(x)+(1-\theta)f(y)$)

$\longleftrightarrow$ $c \le x$, וזה נכון בהנחה ③

- בהנחה ②, אם $\theta x+(1-\theta)y = Max(\theta x+(1-\theta)y, c)$, אז:

$$f(\theta x+(1-\theta)y) = \theta x+(1-\theta)y \underset{x<c}{<} \theta c+(1-\theta)y = \theta f(x)+(1-\theta)f(y)$$

אחרת, $Max(\theta x+(1-\theta)y, c) = c$. לכן:

$$f(\theta x+(1-\theta)y) \le \theta f(x)+(1-\theta)f(y) \longleftrightarrow c \le \theta c+(1-\theta)y \longleftrightarrow \theta c \le \theta y$$

(אם $\theta=0$ אז: $f(\theta x+(1-\theta)y) = f(y) = \theta f(x)+(1-\theta)f(y)$)

$\longleftrightarrow$ $c \le y$, וזה נכון בהנחה ②

לכן בכל מקרה f קמורה מהצורה, כנדרש.

5.

**הפרכה:** ניקח $x=0$, $y=\pi$, $\theta=\frac{1}{2}$, ונקבל:

$$f(\theta x+(1-\theta)y) = \sin\left(\tfrac{1}{2}\pi\right) = 1 > 0 = \tfrac{1}{2}\sin(\pi) = \theta f(x)+(1-\theta)f(y)$$

לכן f לא קמורה.