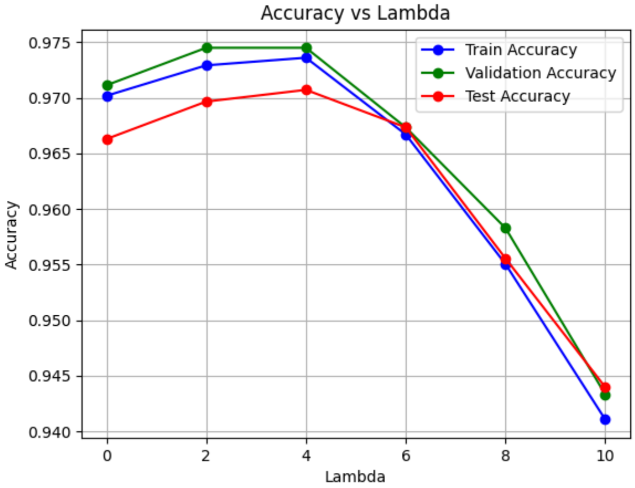
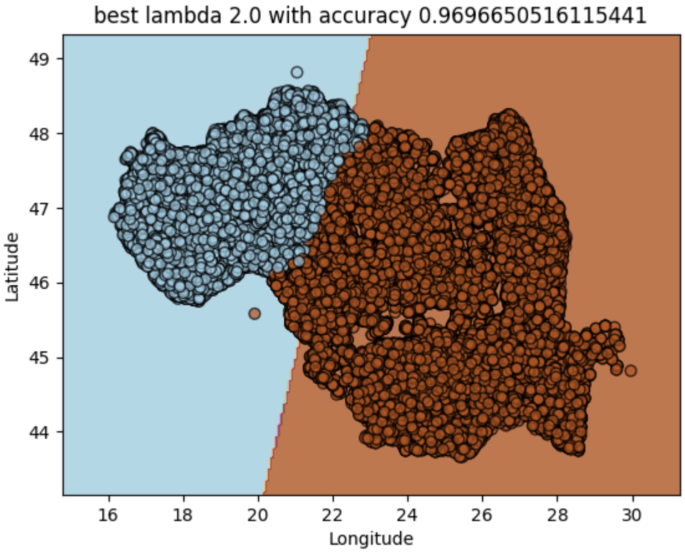
Practical part

3.2

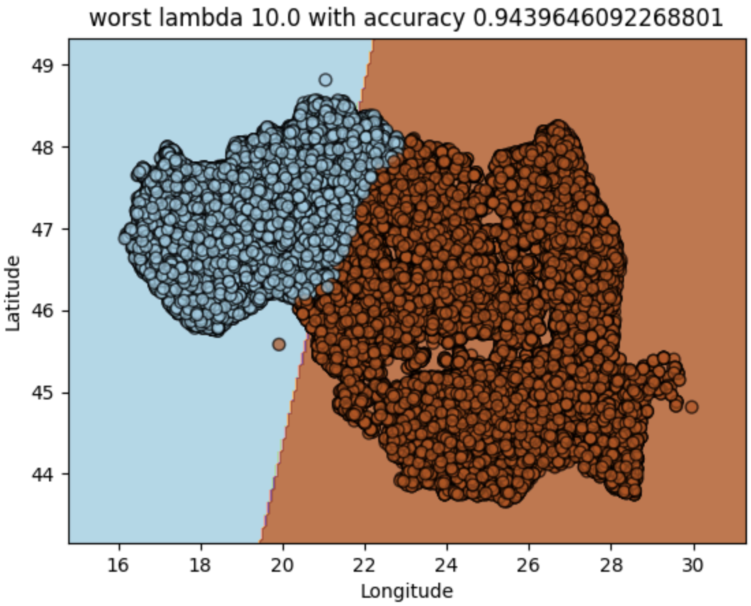


The model that receives the best accuracy is lambda=2 with a rate of 0.9745048461862621 on the validation set. it’s clear from the graph that all models are changing very small in response to the change in lambda.

Best model:



Worst model:



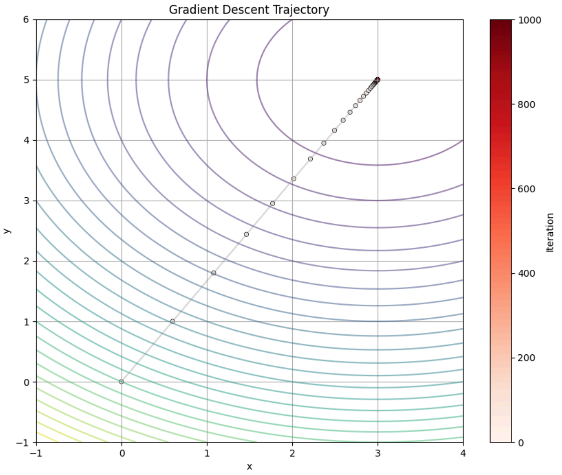
In this case the regulation of lambda makes a little visual difference due to two different factors.

1. The dimension of the created vector is 2 (not including bias), thus creating a small norm.
2. The data is well spread and unnoisy so regulation makes a small difference there are not so many outliers to account and accommodate for.

Gra

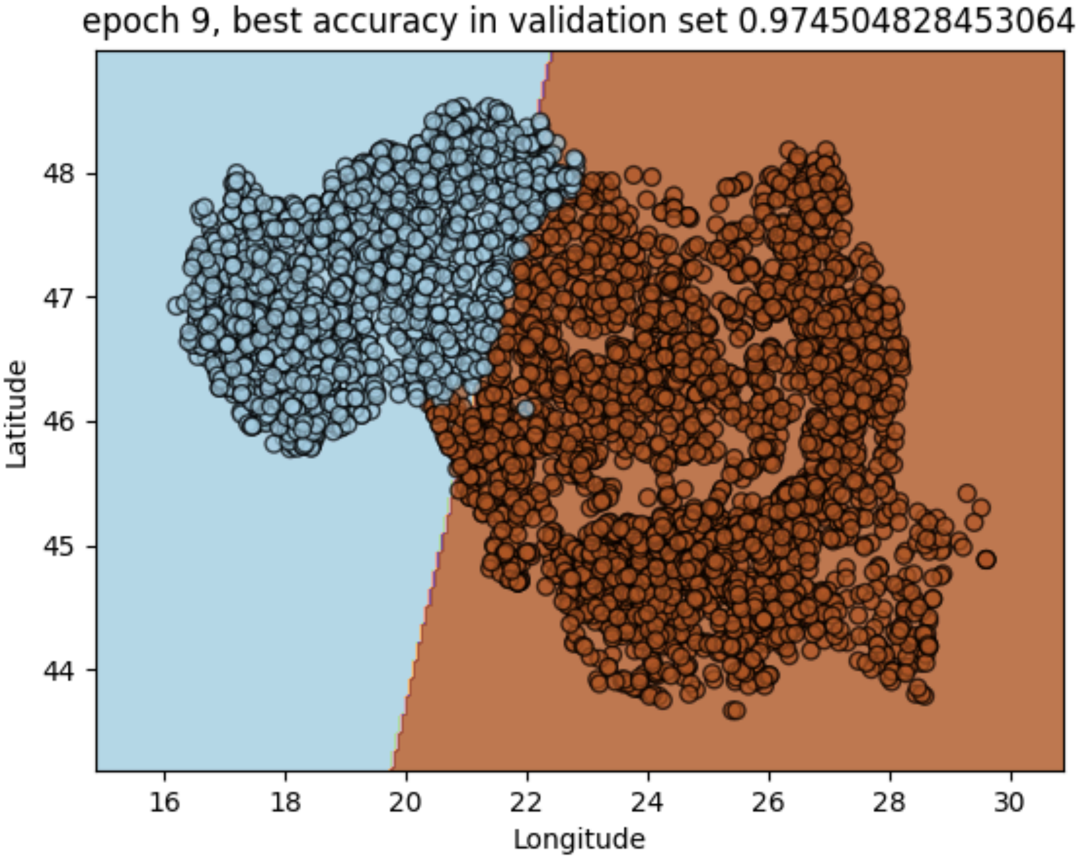
4.1 Gradiant descent with NP

Reached a very close point x=2.999999999999999, y=4.999999999999998, f(x,y)=3.944304526105059e-30. It reaches by the 159 iteration and stays around this area.

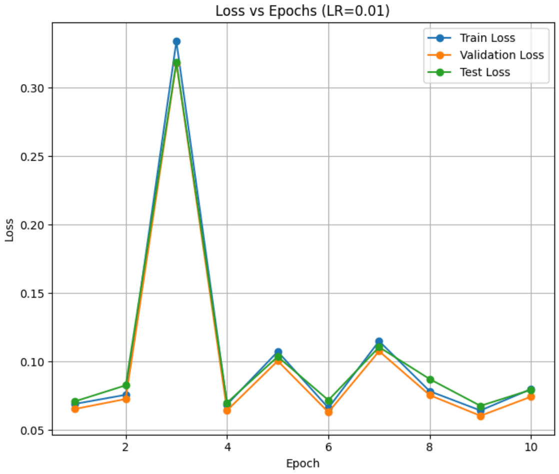


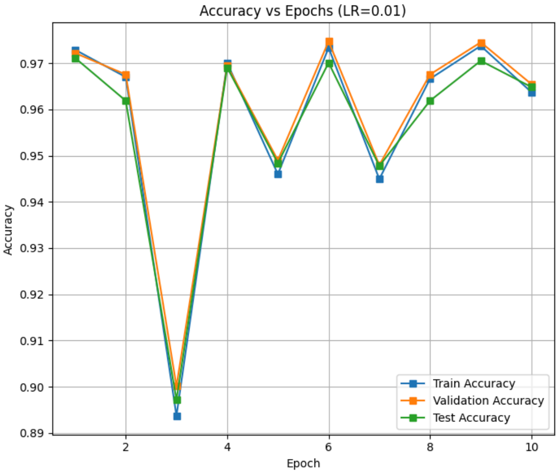
6.3 Logistic Regression the binary case

Best results derived from the eta=0.01 learner with 0.9745 accuracy on validation, and 0.9705 on the
Here is it’s plot:



This is the loss function on all the sets.

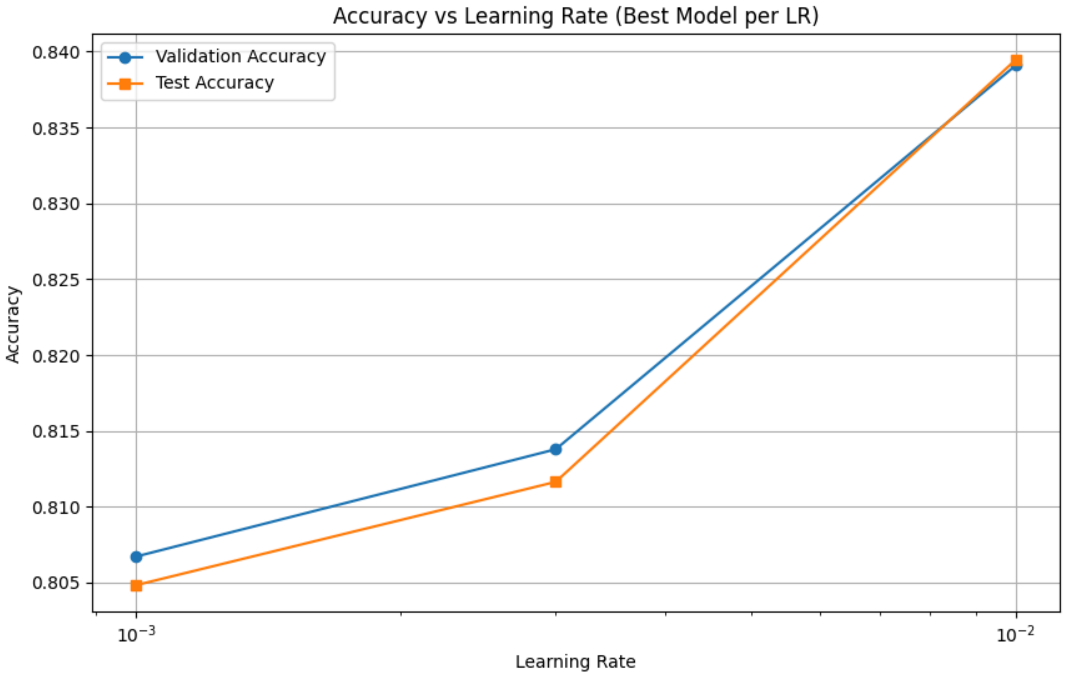




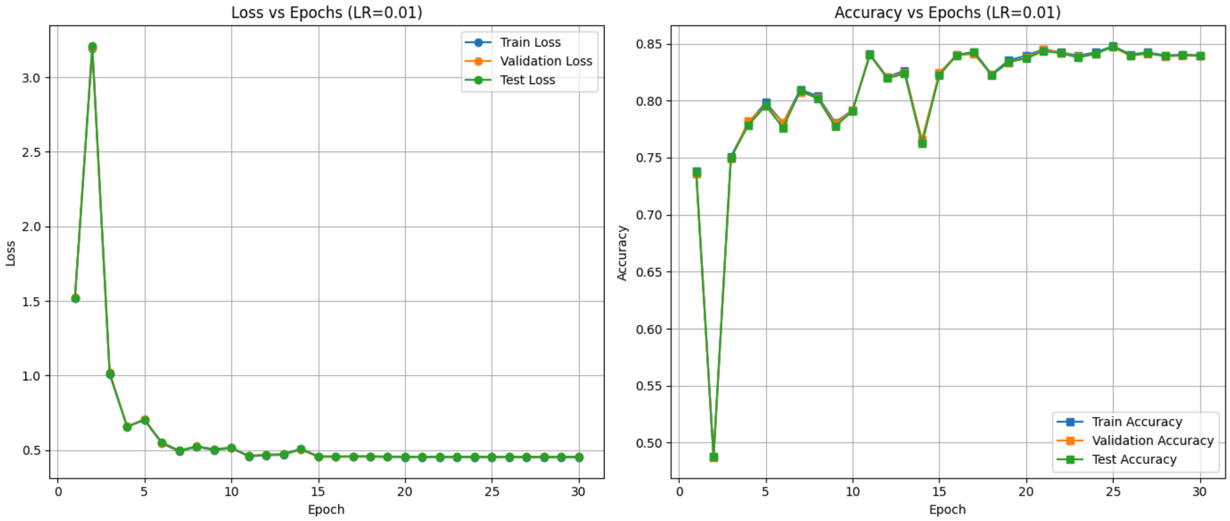
It’s clear that there is a good corelation between success on the validation and on the test set. While on the training set the success is more limited - probably due to it’s size.
It’s also clear that the SGD did randomly ‘walk’ in the wrong direction during the 3rd epoch, as could happened in this algorithm,
Finally, the algorithm worked just as well as the Ridge regression on the final result of it’s ability to guess the states.

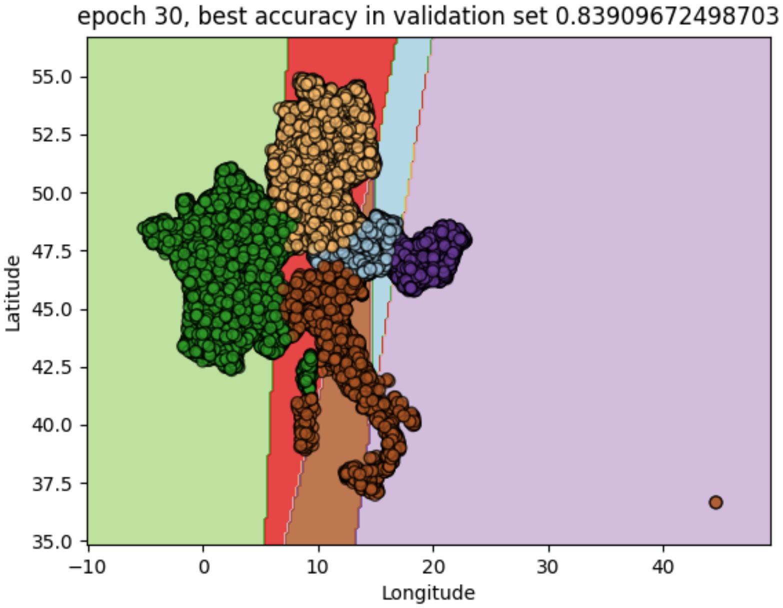
6.4 Logistic Regression Multi-Class

1. The Accuracy is better for the larger step consistently. Probably due to the decay, the step is getting two small by with the smaller learning rates.



2. The model generalizes very well from the training set, you can see a strong correlation between all the sets.

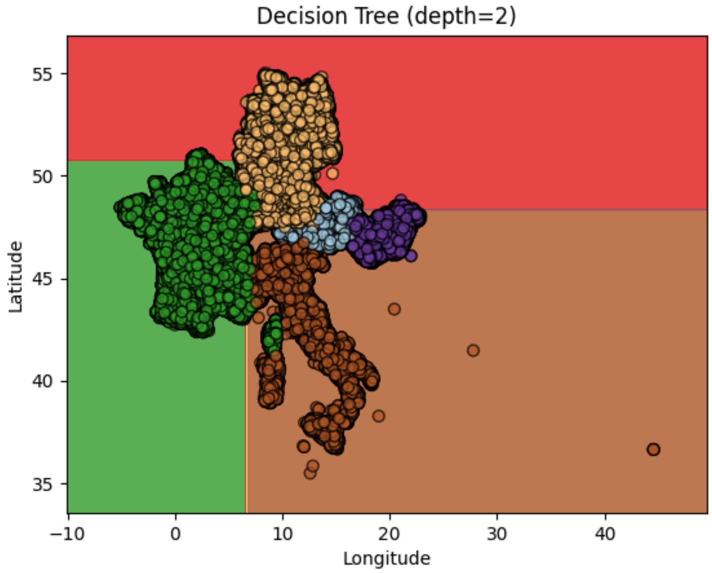




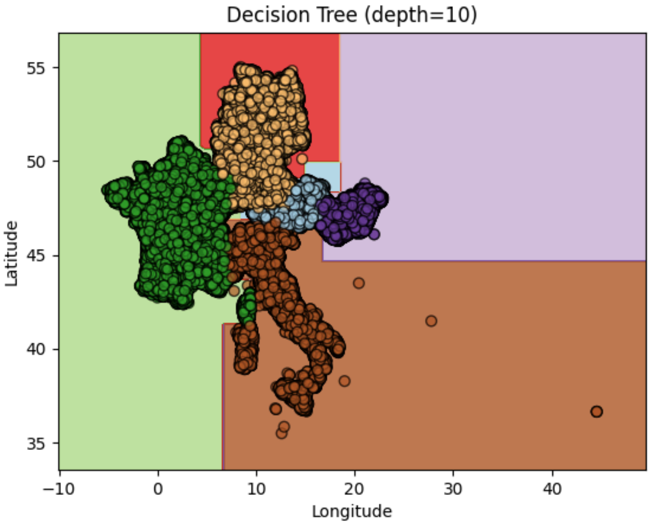
Disclosure

I mostly coded this exercise myself,I used AI mainly for plotting and printing.

3. For the depth 2 tree we get a 0.7502 test accuracy. This of-course is much less suitable the the linear regression discussed on the lats question



4. As for depth 10 we get a much better and more preferable results that are much better at predicting the states. This model is better then all models showed this far, first it's very simple and fast to train, the accuracy is best (Test Accuracy: 0.9969). And as discussed in exercise 2, this model fits very well to axis aligned problems.


Decision Tree (depth=10)
55
50
45
40
35
Latitude
−10
0
10
20
30
40
Longitude

IML - Ex-3

Omri Melcer

December 15, 2025

1. Let $f(x) = ax + b$ we will show that f is convex. Let $x, y \in \mathbb{R}, t \in [0, 1]$.

$$f\left(tx + (1-t)\,y\right) = atx + a(1-t)y + b = atx + a(1-t)y + tb + (1-t)b = t*(f(x)) + (1-t)f(y)$$

2. We will prove that $f(x) = ax^2 + bx + c$ is convex $\iff a \geq 0$:

   (a) $f(x)$ is convex $\implies a \geq 0$. we will take $x = 1$ $y = -1$ and $t = \frac{1}{2}$then we know that

$$f(tx+(1-t)y) = f(0) = c \leq \frac{1}{2}f(-1) + \frac{1}{2}f(1) = \frac{a1^2 - bc + c + ac^2 + bc + c}{2} = a + c \implies 0 \leq a$$

   (b) $a \geq 0 \implies f(x)$ is convex Let $x, y \in \mathbb{R}$and $t \in [0, 1]$. we will show that

$$\Delta = t(f(x) + (1-t)f(y) - [f(tx + (1-t)y)] \geq 0$$

$$f(tx + (1-t)y) = a(tx + (1-t)y)^2 + b(tx + (1-t)y) + c$$

   like in section 1 the $b(tx + (1-t)y) + c$ will be cancelled from both side in Δ. Thus we get

$$\Delta = \left[at^2x^2 - 2at(t-1)xy + a(1-t)^2y^2\right] - tax^2 - (1-t)ay^2$$

   we pull out a and group by x^2, y^2, xy

$$\Delta = a\left[x^2\left(t^2 - t\right) + 2xy\left(t^2 - t\right) + y^2 \underbrace{\left[(1-t)^2 - (1-t)\right]}_{=1-2t+t^2-1+t=t^2-t}\right] = \underbrace{a}_{\geq 0} * \underbrace{t}_{\geq 0} * \underbrace{(1-t)}_{\geq 0}\underbrace{(x-y)^2}_{\geq 0} \geq 0$$

3. $f(x) = e^x$ is convex. The most standard way to prove this is by second derivative since $f''(x) = e^x \geq 0$. Since we only use the algebraic definition, Let $x, y \in \mathbb{R}, t \in [0, 1]$.

$$f(tx + (1-t)y) = e^{tx+(1-t)y} = e^{tx} * e^{(1-t)y}$$

   both $e^x = u,\ e^y = v$ are non-negative then by the young inequality:

$$e^{tx} * e^{(1-t)y} = u^t v^{1-t} \leq tu * (1-t)v = tf(x) + (1-t)y$$

4. Let $c \in \mathbb{R} f(x) = \max(x, c) = \begin{cases} c & x \leq c \\ x & else \end{cases}$. Let $x, y \in \mathbb{R}, t \in [0, 1]$.

   (a) case 1: $x \leq c, y \leq c \implies tx + (1-t)y \leq c \implies f(tx + (1-t)y) = c$ thus:
   $$t(f(x)) + (1-t)f(y) = c \geq c$$

   (b) case 2: $x, y \geq c$ this is a linear function which we already proved is convex.

   (c) case 3: WLG $x \leq c$ and $y \geq c$ thus $f(x) = c$ $f(y) = y$.
   $$tf(x) + (1-t)f(y) = tc + (1-t)y$$
   but also
   $$tc + (1-t)y \geq tc + (1-t)c = c$$
   Thus
   $$\max(c, tc + (1-t)y) \leq tf(x) + (1-t)f(y)$$

5. We will show a counter example to $f(x) = cos(x)$ with $t = \frac{1}{2}$ and $x = -\frac{\pi}{2}$ and $y = \frac{\pi}{2}$ thus on the left side we get $cos(0) = 1$ and on the right side we get
$$\frac{cos(-\frac{\pi}{2}) + cos(\frac{\pi}{2})}{2} = 0 \ngeq 1$$