REPORT - OHAD DANIEL

Theoretical Part - Convexity

1. תהי $f:\mathbb{R}^d\to\mathbb{R}^d$ $f(x)=ax+b$. תהי $\theta\in[0,1]$.

נבחר בנקודות $x,y\in\mathbb{R}^d$ ונציב בנוסחה:

$$f(\theta x+(1-\theta)y)=a(\theta x+(1-\theta)y)+b$$
$$=a\theta x+ay-\theta ay+b=*$$

כעת נציב בנוסחה השנייה של אי השוויון:

$$\theta f(x)+(1-\theta)f(y)=\theta ax+\theta b+ay+b-\theta ay-\theta b$$
$$=\theta ax+ay-\theta ay+b=**$$

אני נשים לב כי לכל x,y $**=*$ ובפרט מקיים את $*\le**$ כנדרש.

2. תהי $y,x\in\mathbb{R}^d$. ותהי $\theta\in[0,1]$.

נבדוק האם הפונקציה קמורה לפי קריטריון קמירות נגזרת שנייה:

$$f(x)=ax^2+bx+c$$
$$f'(x)=2ax+b$$
$$f''(x)=2a.$$

אם״ם $a\ge0$ אז $f''(x)\ge0$ כלומר $f(x)$ קמורה.

3. נוכיח לפי קריטריון חיוביות הנגזרת השנייה: תהי $x\in\mathbb{R}$, ו $f:\mathbb{R}\to\mathbb{R}$ $f(x)=e^x$.

$$f(x)=e^x$$
$$f'(x)=e^x$$
$$f''(x)=e^x$$

כעת נשים לב כי לכל $x\in\mathbb{R}$ $e^x>0$ ובפרט $f''(x)\ge0$ לכן $f(x)$ קמורה.

4. $f(x)=\max(x,c)$ היא מקסימום של מספר סופי (2) של פונקציות קמורות $f(x)=ax+b$ כאשר $a=1,b=0$ כפי שראינו. ו $f(x)=c$ הפונקציה הקבועה, לכן כיוון שמקסימום של מספר סופי של פונקציות קמורות היא קמורה נקבל ש $f(x)$ קמורה לכל $c\in\mathbb{R}$.

5. $f(x)=\cos(x)$ איננה קמורה. נראה לפי קריטריון נגזרת שנייה:

$$f(x)=\cos(x)$$
$$f'(x)=-\sin(x)$$
$$f''(x)=-\cos(x)$$

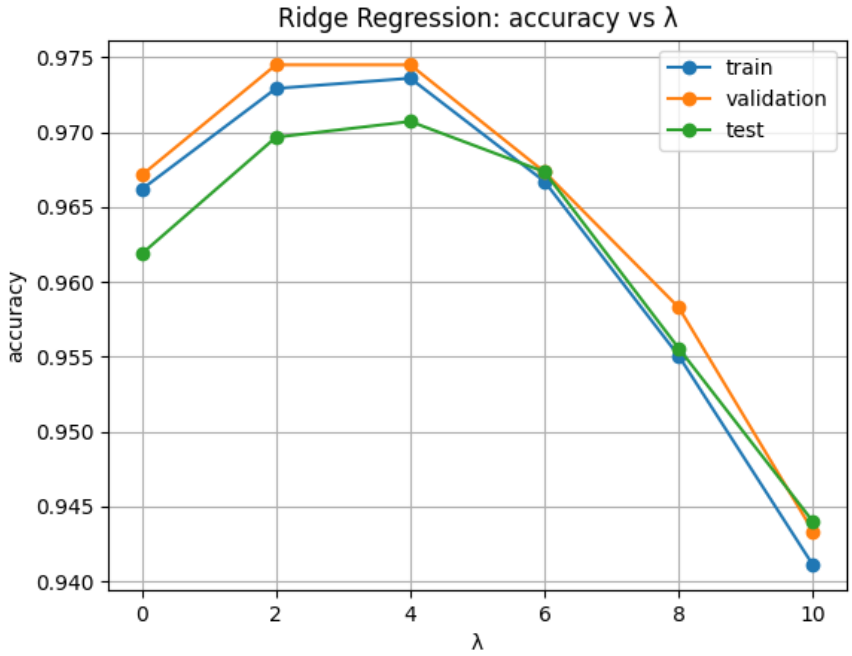
כעת נבחר בקטעים $[-\frac{\pi}{2}+2\pi k,\frac{\pi}{2}+2\pi k]$ עבור $k\in\mathbb{R}$.

נשים לב כי בקטעים אלה $-\cos x<0$ כלומר הפונקציה קעורה.

ולכן הפונקציה $\cos(x)$ איננה קמורה.

Ridge Regression - Analytical Solution

Effect of the Ridge Regularization Parameter (λ) on Model Accuracy:

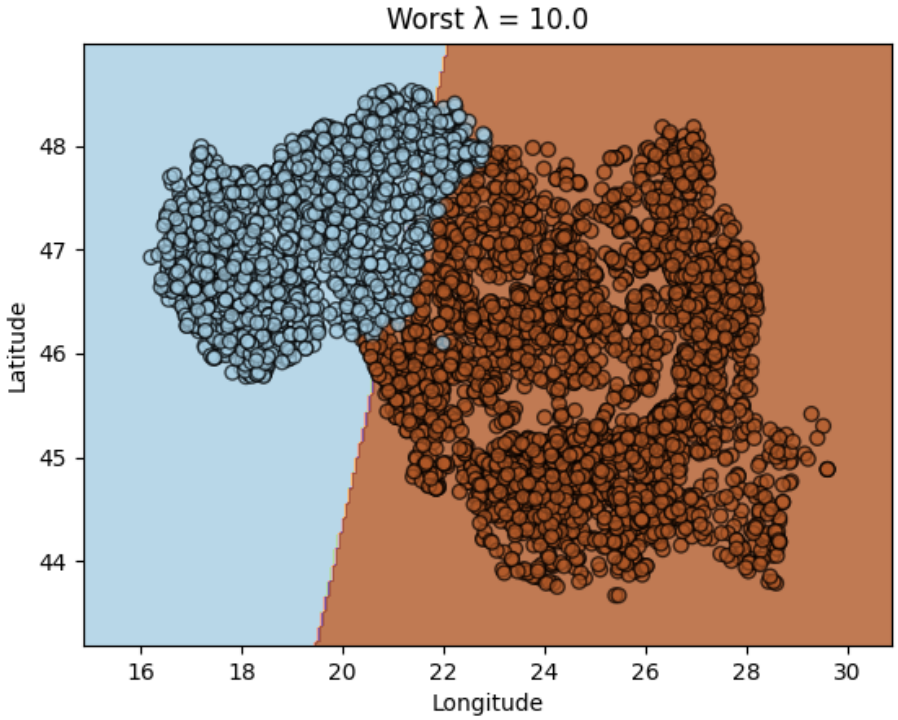


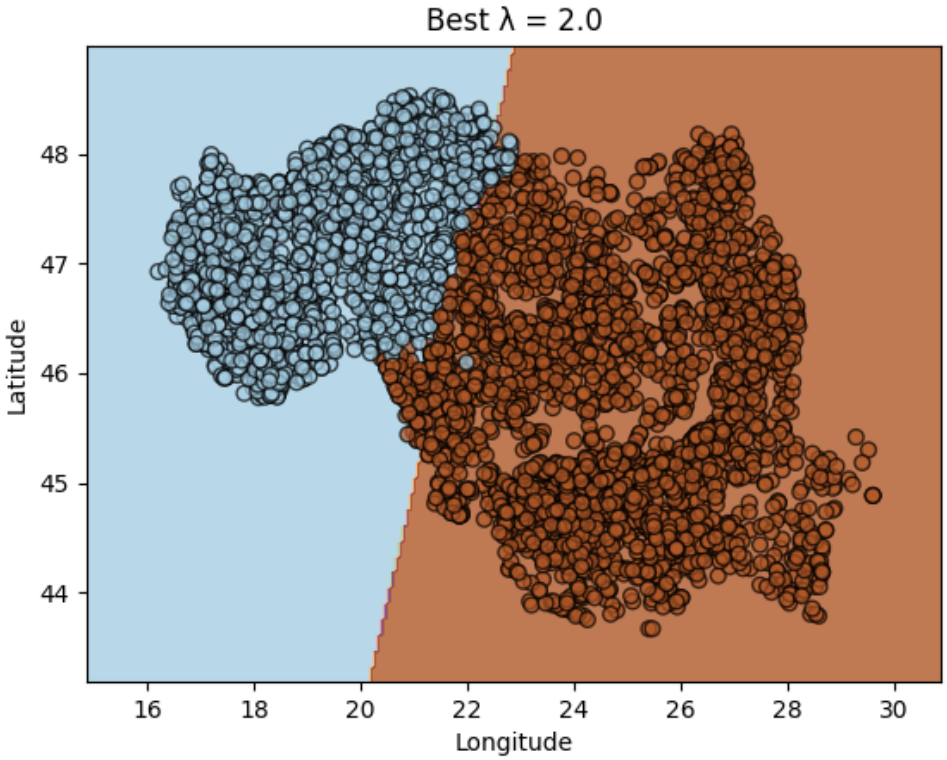
We trained a ridge regression classifier with. $\lambda \in \{0,2,4,6,8,10\}$and plotted the train, validation and test accuracies as a function of λ.

The plot shows that for very small regularization $(\lambda = 0)$ the model already achieves relatively high accuracy on all sets, but the validation and test accuracies improve slightly when increasing λ to 2–4. For larger values of$\lambda\,(6,8,10)$ all three accuracies start to decrease, indicating that the model becomes too constrained (underfitting).

The best validation performance is obtained at λ = 2, with validation accuracy ≈ 0.975. The corresponding test accuracy of this best model is approximately 0.97.

Visualization of Decision Boundaries for Different λ Values:



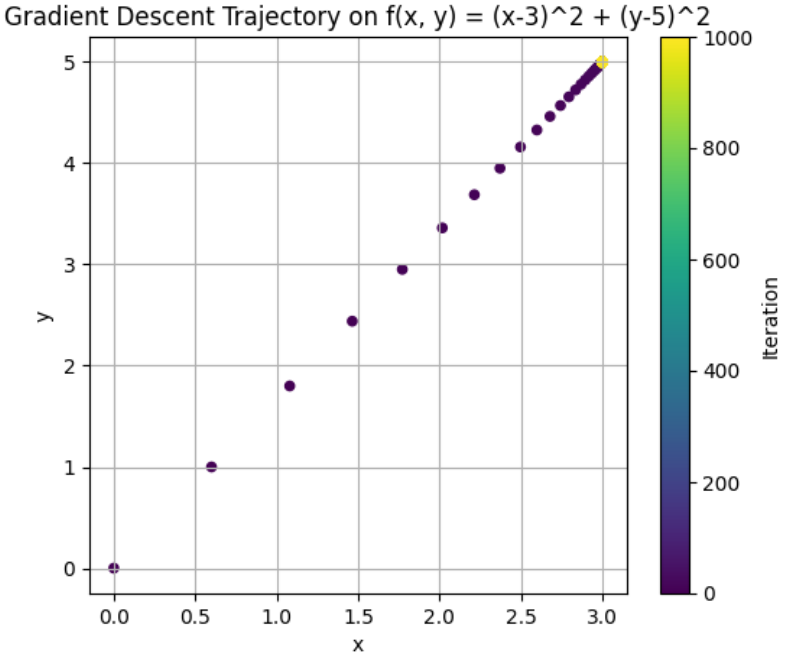


Using the visualization helper, we plotted the prediction space (decision boundaries) of the ridge classifier for the best $\lambda(\lambda = 2)$ and the worst $\lambda(\lambda = 10)$, where “best” and “worst” are chosen according to the validation accuracy. In both plots, the background color represents the class predicted by the model at each location in the (longitude, latitude) plane, and the test cities are shown as points colored by their true country label.

For $\lambda = 2$, the linear decision boundary aligns well with the two clusters of points: most cities of each class lie on the correct side of the boundary, so the background color around each cluster matches the point colors, reflecting the high test accuracy (~0.97). For$\lambda = 10$, the boundary is noticeably less well positioned relative to the clusters: more points fall in regions where the background color does not match their true label, corresponding to the lower accuracy.

This illustrates how λ affects the algorithm: small-to-moderate regularization $(e.g., \lambda = 2)$ stabilizes the model and improves generalization, while overly large λ forces the weights towards zero, making the classifier too simple (underfitting) and degrading its ability to separate the two countries in the feature space.

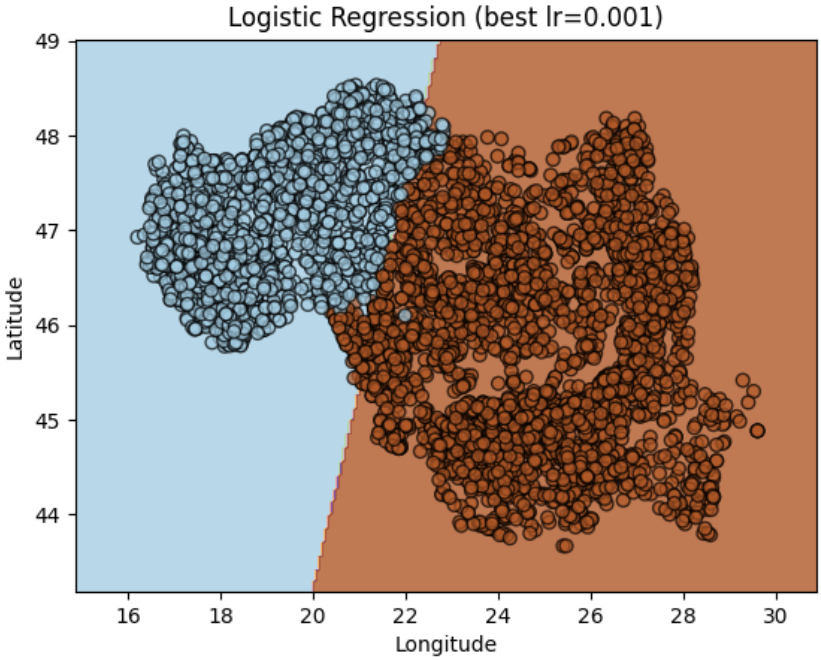
Gradient Descent in NumPy



Gradient descent with step size 0.1 for 1000 iterations, starting at (0,0), converges to approximately (x,y) = (3,5), which is the minimum of $f(x, y) = (x - 3)^2 + (y - 5)^2$. .

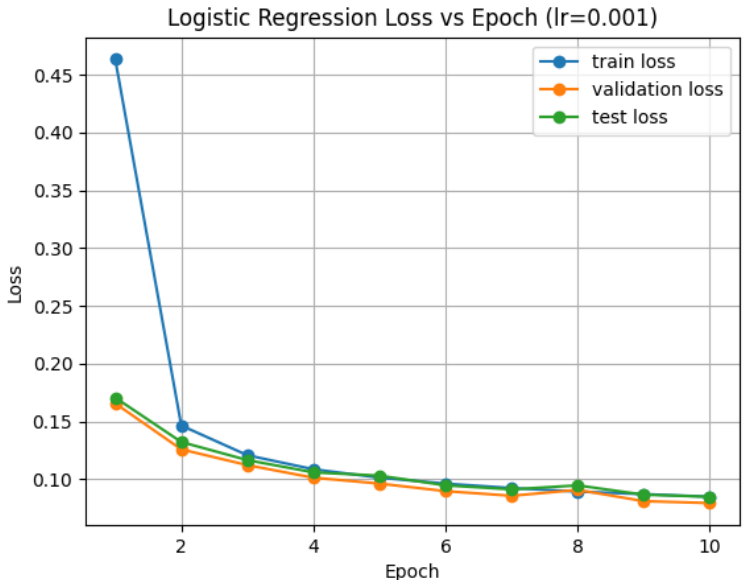
Logistic Regression - Stochastic Gradient Descent

Choosing and Visualizing the Best Logistic Regression Model:



The best model was obtained with learning rate $\eta = 0.001$.

Loss Curves and Generalization Behaviour



Loss Curves and Generalization Behaviour

For the best model $(\eta = 0.001)$ we plotted the training, validation and test losses over the 10 epochs.

All three losses decrease rapidly during the first few epochs and then continue to decrease more slowly, stabilizing around a similar value (≈0.08–0.09). The curves for train, validation and test are close to each other throughout training, without a large gap where the training loss continues to drop while validation/test losses increase.

This pattern indicates that the model generalizes well: it learns a good decision boundary from the training data and maintains similarly low loss on the unseen validation and test sets (no strong overfitting or underfitting is observed).

Comparison to Ridge Regression (Section 3.2)

In Section 3.2 we trained a ridge regression classifier on the same data, selecting the best λ according to validation accuracy ($\lambda = 2$). That model also achieved high test accuracy and a linear decision boundary that separates the two countries well.

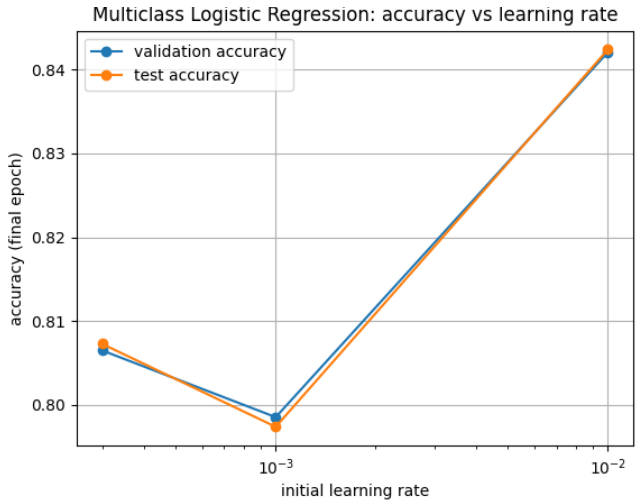
Logistic regression and ridge regression are both linear models, but they optimize different objectives:

• Ridge regression minimizes squared error with anL_2 penalty, treating the problem as regression and then thresholding the output to obtain class labels.
• Logistic regression directly optimizes the cross-entropy classification loss (via softmax + CrossEntropyLoss), which is more naturally aligned with a probabilistic binary classification task.

Because of this, logistic regression is often slightly better suited to classification, while ridge regression can perform similarly well when the classes are nearly linearly separable, as in this dataset. In our experiment, both methods produced very similar linear boundaries and high test accuracy; logistic regression achieves this via gradient-based optimization of a classification loss, whereas ridge regression uses a closed-form solution for a regularized least-squares problem.

Multi-Class Case

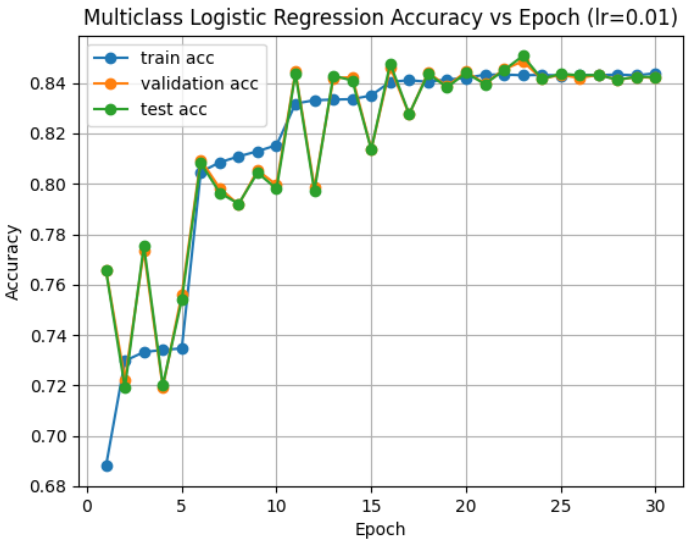
Q1 – Effect of the Initial Learning Rate on Multiclass Logistic Regression

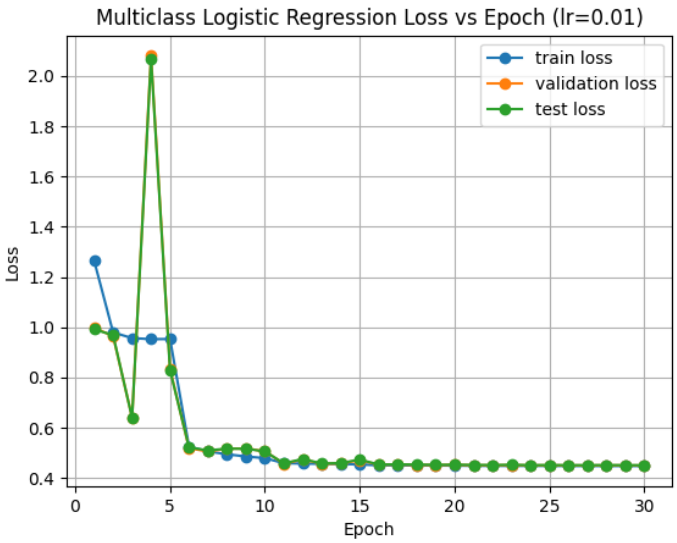


We trained multiclass logistic regression models with initial learning rates $\eta \in \{0.01, 0.001, 0.0003\}$. using 30 epochs, batch size 32 and learning-rate decay by a factorof 0.3 every 5 epochs. For each learning rate we recorded the final validation and test accuracies and plotted them as a function of the learning rate

The plot shows that small learning rates ($\eta = 0.001, 0.0003$) reach validation accuracies of about 0.80–0.81, while ($\eta = 0.001, 0.0003$) achieves a noticeably higher validation accuracy of about 0.84. According to the validation set, the best model is therefore the one with initial learning .rate $\eta = 0.01$, whose corresponding test accuracy at the end of training is approximately 0.842

Q2 – Training Dynamics and Generalization of the Best Multiclass Logistic Regression Model



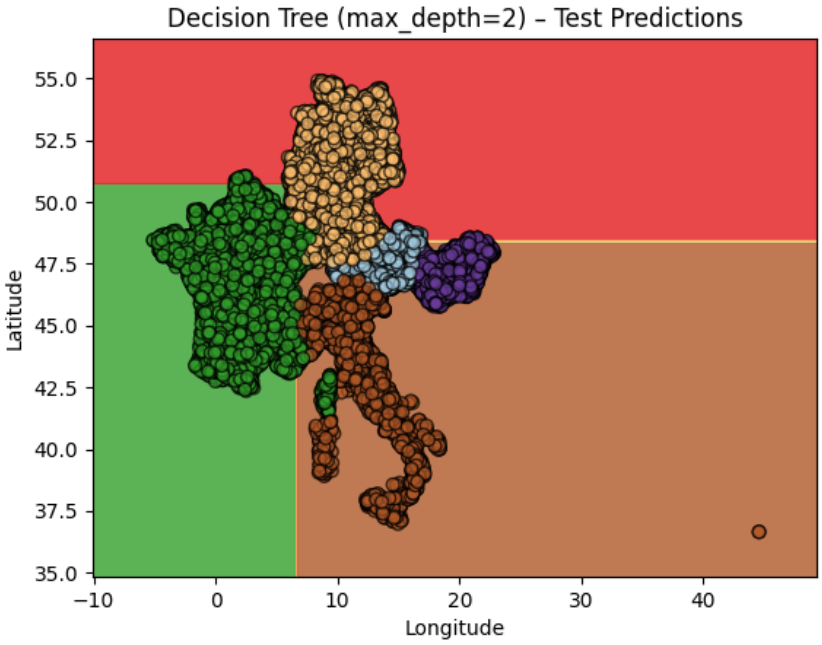


For the best model (initial learning rate $\eta = 0.01$ with decay by 0.3 every 5 epochs), we plotted training, validation and test losses and accuracies over 30 epochs

The loss curves show that after a somewhat noisy start (including a temporary spike in validation/test loss when the learning rate is still high), all three losses rapidly decrease once the learning rate decays and then stabilize around a similar value (~0.45–0.47). The accuracy curves rise from about 0.70 to roughly 0.84–0.85, with the training, validation and test accuracies remaining very close to one another throughout training

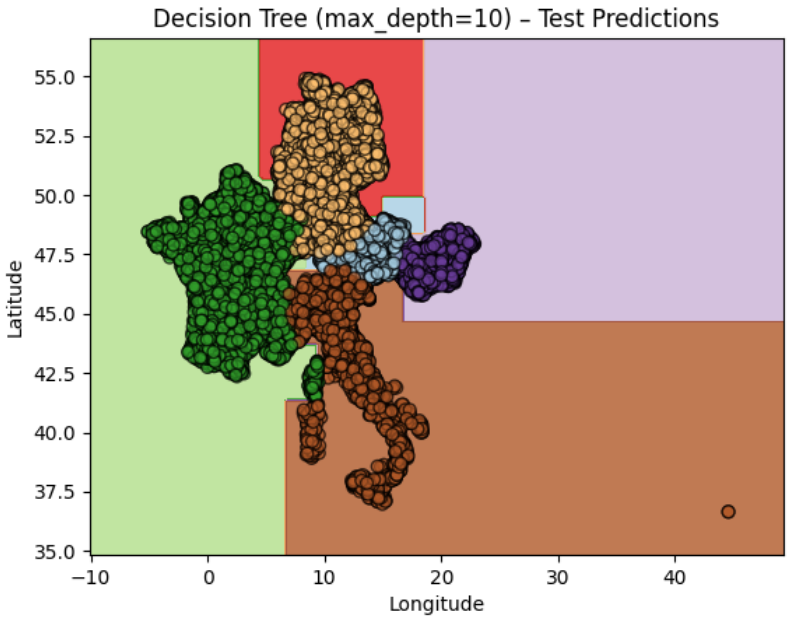
Because the validation and test curves closely track the training curves, without a large gap or late increase in validation/test loss, this model generalizes well: it learns a good decision boundary on the training data while maintaining comparable performance on unseen data

Q3 – Decision Tree (max depth = 2) vs. Logistic Regression



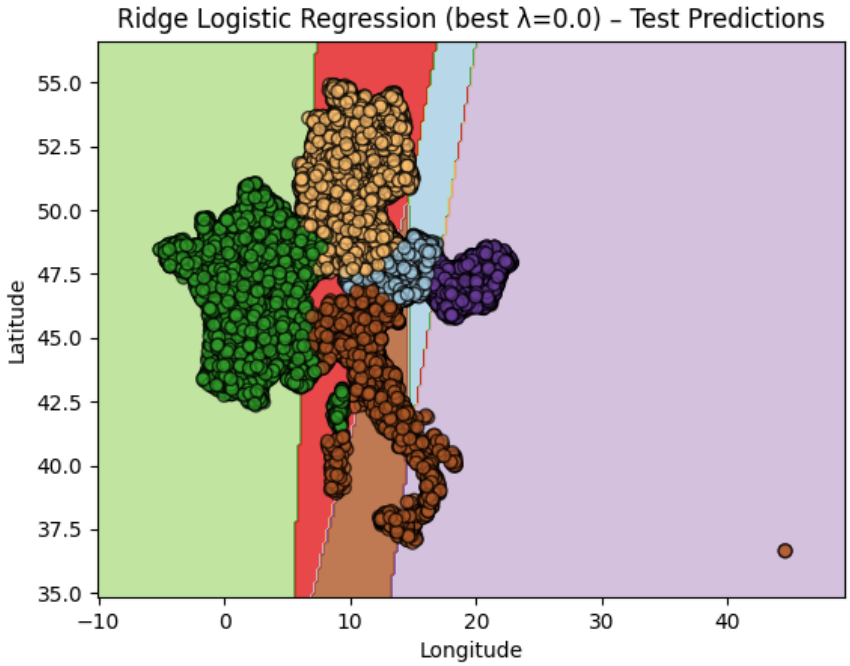
We trained a DecisionTreeClassifier with max_depth = 2 on the multiclass dataset. Its accuracies were: train accuracy 0.751, validation accuracy 0.750, and test accuracy 0.750. The decision-boundary plot shows very coarse axis-aligned rectangular regions. The tree splits the plane into a few large blocks, so many cities near the true borders between countries are misclassified. Compared to the multiclass logistic regression model from Q2 (test accuracy about 0.842), the shallow tree clearly underfits: it has higher bias, lower accuracy, and much less precise boundaries. Therefore, the logistic regression model is more suitable for this task than the depth-2 decision tree

Q4 – Deep Decision Tree (max_depth = 10) and Comparison to Logistic Regression



We then trained a deeper DecisionTreeClassifier with max_depth = 10. Its accuracies were: train accuracy 0.997, validation accuracy 0.996, and test accuracy 0.997. The corresponding decision-boundary plot shows many small, irregular regions that closely follow the clusters of points. The model is much more flexible than the depth-2 tree and is able to almost perfectly separate the classes. In contrast to Q3, this deep tree outperforms the logistic regression model from Q2 (test accuracy about 0.842) by a large margin, while still having very similar train, validation and test accuracies. That means the tree has low bias and, in this dataset, its higher variance does not hurt generalization. Therefore, for this task, the max_depth = 10 decision tree is more suitable than the logistic regression model, and our conclusion changes compared to Q3

Q5 – Effect of Ridge Regularization on Multiclass Logistic Regression



We repeated the multiclass logistic regression experiment, adding an L2 (ridge) penalty with lambda in {0, 2, 4, 6, 8, 10}, using the same training procedure as in Q2. For each lambda we measured validation and test accuracies at the end of training. The best model according to the validation set was obtained with lambda = 0.0, i.e., no ridge regularization: best lambda 0.0, validation accuracy 0.842, and test accuracy 0.842. For larger lambda values the validation accuracy dropped significantly (for example around 0.73 for lambda = 2 and down to about 0.66 for lambda = 10), indicating underfitting due to overly strong regularization. The decision-boundary plot for the best ridge model (lambda = 0) is essentially the same as the logistic regression model from Q2, since the optimal lambda is zero. In this problem, adding ridge regularization does not improve performance; in fact, it hurts when lambda is too large. The unregularized logistic regression model remains the best linear model, and the deep decision tree from Q4 is still the strongest overall classifier in terms of accuracy

Q6

BROWN -ITALY
GREEN - FRANCE
YELLOW - GERMANY
LIGHT BLUE - SWITZERLAND
PURPLE - AUSTRIA