Understanding Regularisation in Logistic Regression:

Balancing Underfitting and Overfitting
Student Name: Srikanth Pasunoori
Student ID:24090495

---

1. Introduction

Logistic Regression is one of the most widely used classification algorithms in machine learning. Despite its simplicity, it remains a strong baseline in many real-world applications due to its interpretability, efficiency, and solid theoretical foundations. However, like all machine learning models, Logistic Regression is vulnerable to both underfitting and overfitting depending on how it is configured.

A key mechanism used to control model complexity in Logistic Regression is **regularisation**. Regularisation constrains the size of model parameters, helping the model generalise better to unseen data. Choosing an appropriate regularisation strength is therefore essential for achieving good performance.

The goal of this tutorial is to demonstrate **how regularisation strength affects decision boundaries and predictive performance in Logistic Regression**. Using a synthetic two-dimensional dataset, we visualise how changing the regularisation parameter alters model behaviour and discuss the implications for generalisation and ethical model deployment.

---

2. Logistic Regression: intuition

Logistic Regression is a linear classification model that estimates the probability of an input belonging to a particular class. It models a linear combination of input features and passes this combination through a sigmoid function to produce probabilities between 0 and 1. A decision boundary is then formed where this probability crosses a chosen threshold, typically 0.5.

Although the model is linear in its parameters, Logistic Regression is powerful due to its probabilistic interpretation and ease of training. The simplicity of the linear decision boundary makes the model easy to interpret, which is especially valuable in high-stakes applications such as healthcare or finance.

However, the linear nature of the model means that it can struggle with complex datasets. Without additional constraints, Logistic Regression may fit noise in the data, leading to overly confident predictions. Regularisation is introduced to mitigate this issue by discouraging excessively large parameter values.

---

3. Regularisation and model complexity

Regularisation is a technique used to prevent overfitting by penalising complex models. In Logistic Regression, L2 regularisation is commonly used. This penalty discourages large weights by adding a constraint based on the squared magnitude of model coefficients.

Mathematical Formulation of Logistic Regression with Regularisation

Logistic Regression models the conditional probability of a binary class label $y \in \{0,1\}y$ in as:

$$P(y = 1 \mid x) = \sigma(w^T x + b)$$

where $\sigma(z)=1+e-z1$ is the sigmoid function.

The loss function for Logistic Regression with L2 regularisation is:

$$L(w) = -(1/N)\, \Sigma_i \, [\, y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i\,) \,] + (\lambda/2)\, \|w\|^2$$

where:

- first term = logistic loss
- second term = regularisation penalty
- $\lambda = 1 / C$ in scikit-learn

The gradient update becomes:

$$w \leftarrow w - \eta(\partial w \partial L + \lambda w)$$

This shows **explicitly** how strong regularisation shrinks the weights and simplifies the model.

In scikit-learn, regularisation strength is controlled using the parameter **C**, which is the inverse of regularisation strength:

- **Small C** → strong regularisation
- **Large C** → weak regularisation

When regularisation is strong (small C), model coefficients are heavily constrained, leading to simpler decision boundaries. This reduces variance but increases bias, potentially causing underfitting. When regularisation is weak (large C), the model has more freedom to fit the training data closely, increasing variance and the risk of overfitting.

Understanding this bias–variance trade-off is crucial for effective model development and deployment.

---

4. Experimental setup

To demonstrate the effect of regularisation, a synthetic dataset was generated using the make_moons function. This dataset is non-linearly separable and is well suited for visualising decision boundaries in two dimensions.

The dataset was split into training and testing sets using a 70:30 ratio. Logistic Regression models were trained using multiple values of the regularisation parameter C, spanning several orders of

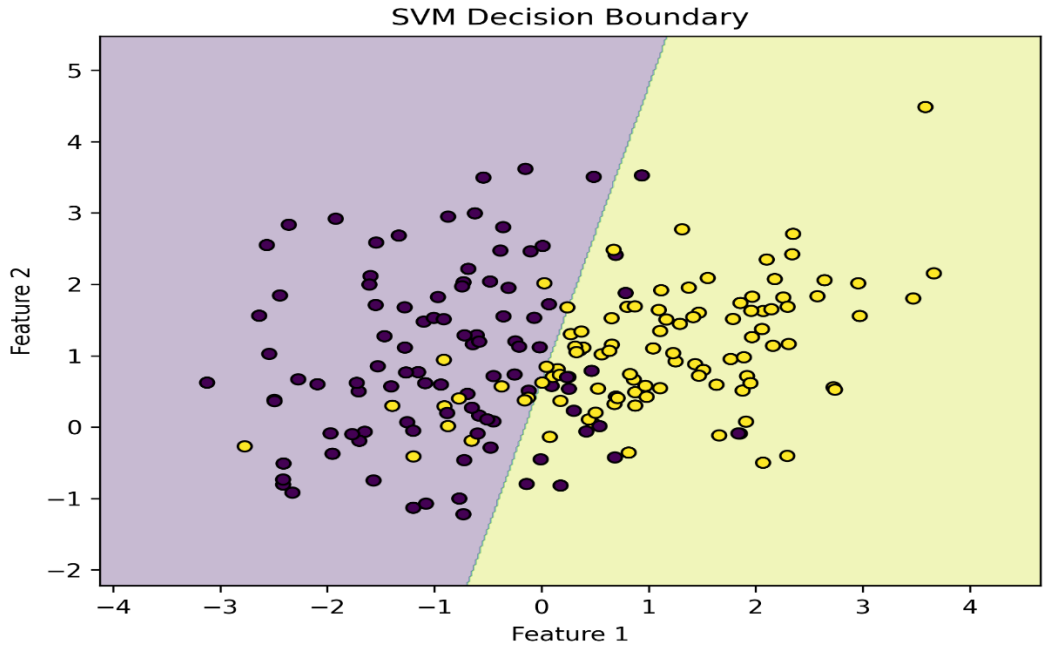
magnitude. For each model, performance was evaluated using classification accuracy on both the training and test datasets.

All experiments were implemented in Python using scikit-learn, ensuring reproducibility and clarity.

---

5. Results and visualisation

Visual inspection of decision boundaries reveals clear differences in model behaviour as regularisation strength changes.

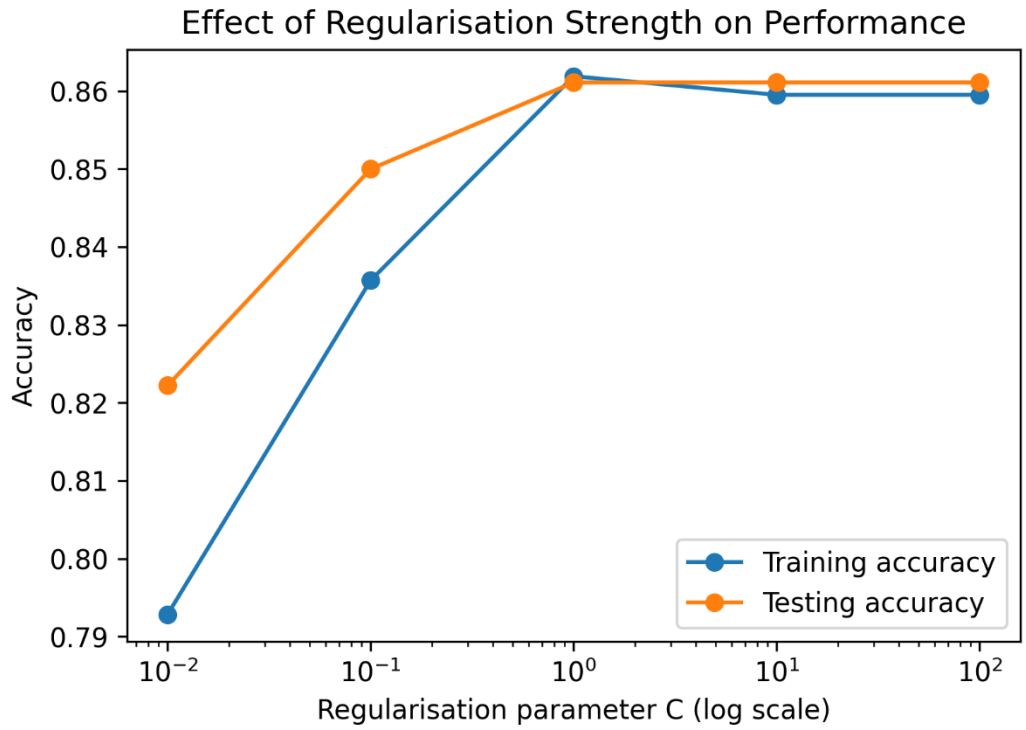


(Figure 1. Linear decision boundary produced by an SVM classifier.
The shaded regions indicate predicted class labels. The diagonal boundary shows how the model separates the two clusters in feature space. This visualisation helps illustrate how linear models behave under different regularisation strengths.)

With **strong regularisation (small C)**, the decision boundary is overly smooth and fails to capture meaningful structure in the data, leading to underfitting. This is reflected in lower training and test accuracy.

As regularisation strength is reduced, the decision boundary becomes more flexible and better aligned with the data. At moderate values of C, the model achieves a balance between bias and variance, producing the best generalisation performance.

With **very weak regularisation (large C)**, the decision boundary becomes highly sensitive to the training data. While training accuracy continues to increase, test accuracy plateaus or decreases, indicating overfitting.



(Figure 2. Training and testing accuracy as a function of the regularisation parameter C (plotted on a logarithmic scale).
Accuracy increases as regularisation becomes weaker, reaching optimal performance at moderate C values. Beyond this point, training accuracy continues to rise while test accuracy plateaus, indicating overfitting.)

These results demonstrate how regularisation directly controls model complexity and highlights the importance of tuning C in practice.

5.1 Decision boundaries at different regularisation strengths (C values)

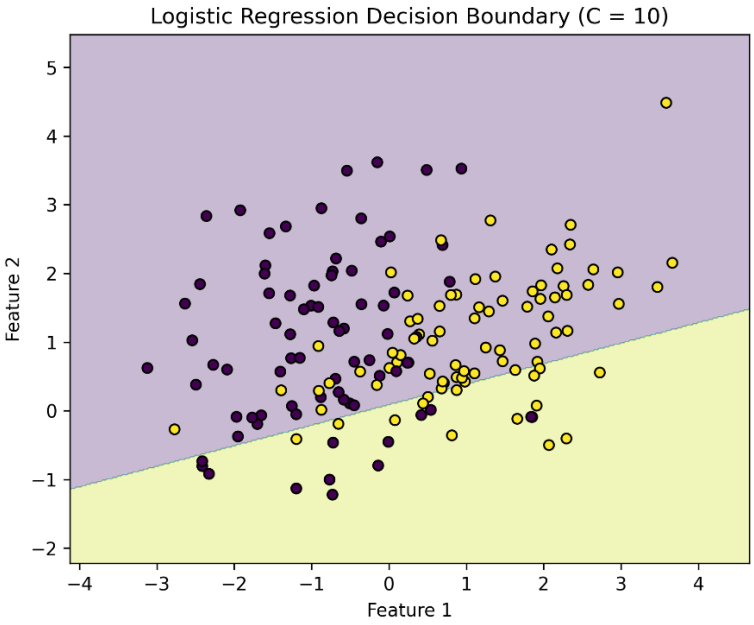


Figure 5.1.1:Logistic Regression Decision Boundary (C=10)

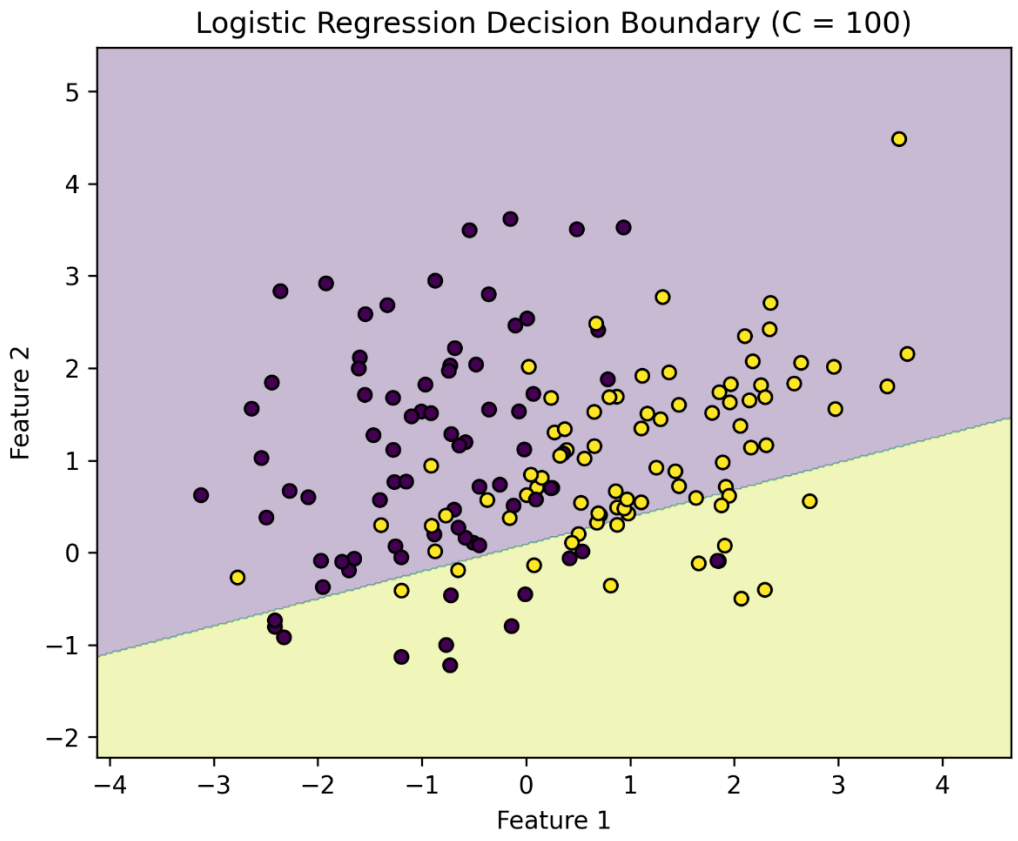


Figure 5.1.2:Logistic Regression Decision Boundary (C=100)

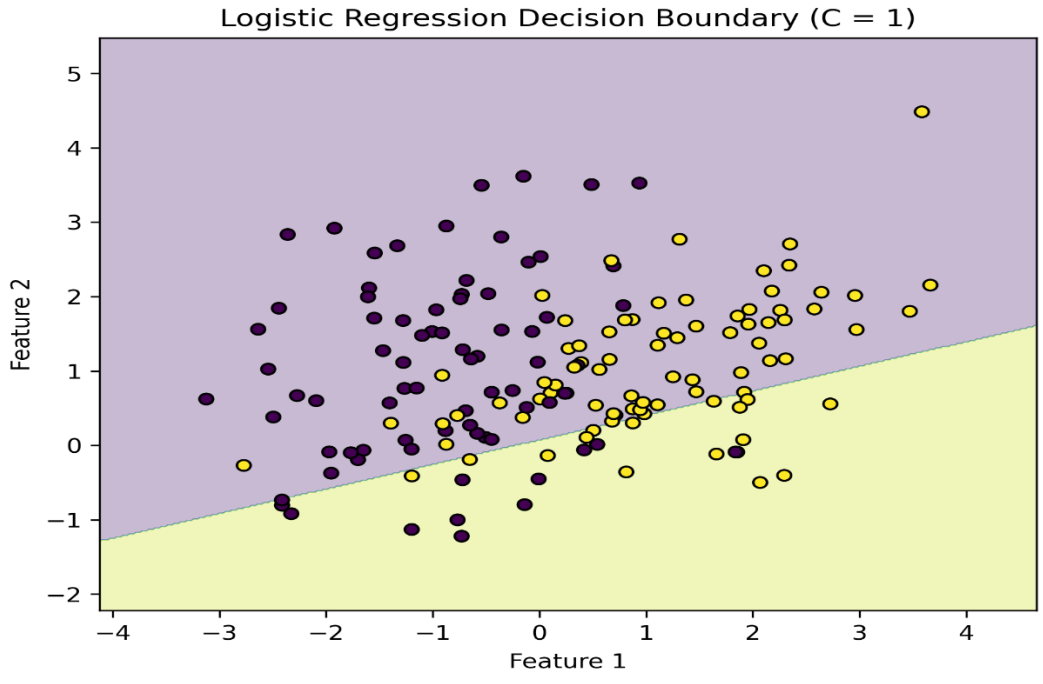


Figure 5.1.3:Logistic Regression Decision Boundary (C=1)

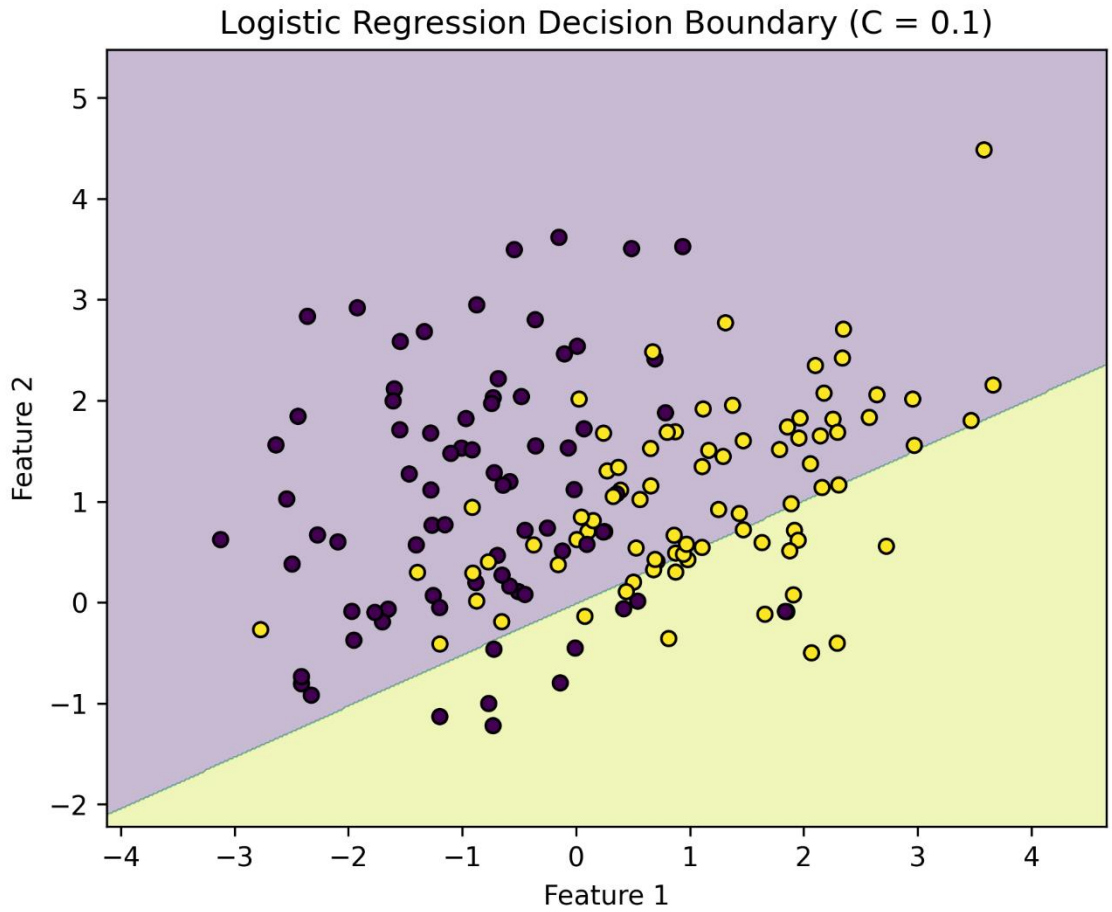


Figure 5.1.4:Logistic Regression Decision Boundary (C=0.1)

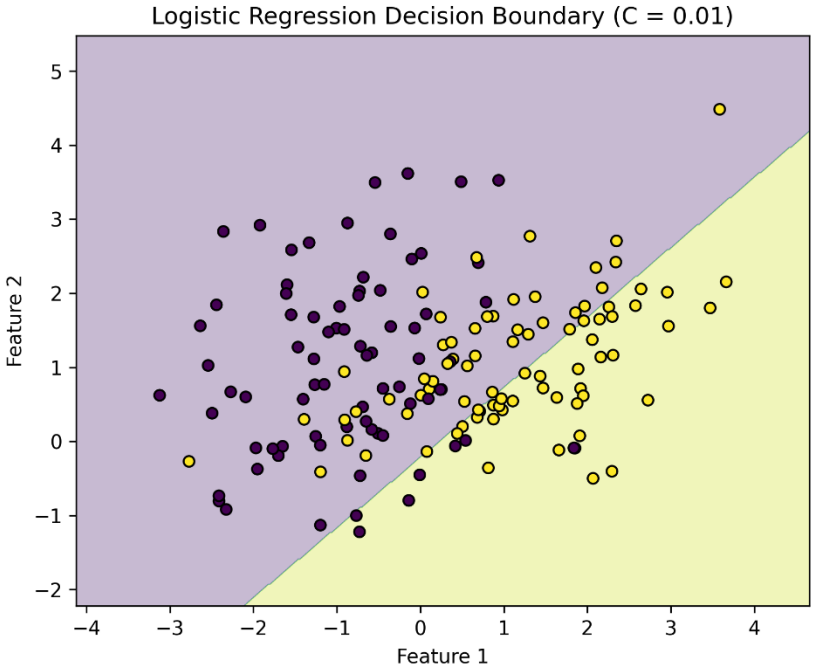


Figure 5.1.5:Logistic Regression Decision Boundary (C=0.01)

5.2: L1 vs L2 Regularisation (Advanced Insight)

While this tutorial focuses on L2 regularisation, it is useful to contrast it with L1 regularisation.

- **L2 (Ridge)** shrinks weights smoothly, encourages small—but non-zero—coefficients, and produces stable solutions.
- **L1 (Lasso)** drives some coefficients **exactly to zero**, performing feature selection.

In high-dimensional settings, L1 can make the model sparse and easier to interpret, while L2 provides better numerical stability. Logistic Regression in scikit-learn supports both, and the choice can strongly affect generalisation.

Summary:

C value	Regularisation strength	Model behaviour	Risk
Very small(0.01)	Very strong	Over simplifies boundary	Underfitting
Small(0.1)	Strong	Smooth boundary	Possible underfit
Medium	Moderate	Best balance	Optimal performance
Large(10-100)	Weak	Very flexible boundary	Overfitting

---

6. Limitations and ethical considerations

Despite its advantages, Logistic Regression has important limitations. Its reliance on linear decision boundaries restricts its ability to model complex patterns without feature engineering. Regularisation cannot overcome fundamental representational limitations of the model.

From an ethical perspective, over-regularisation may lead to systematic underfitting, disproportionately harming minority groups in imbalanced datasets. Conversely, insufficient regularisation can cause models to learn spurious correlations, leading to unreliable or unfair predictions.

Careful regularisation tuning is therefore not only a technical concern but also an ethical responsibility, especially in real-world decision systems.

---

7. Conclusion

This tutorial has demonstrated how regularisation strength influences decision boundaries and performance in Logistic Regression. Through visualisation and empirical analysis, we showed how regularisation governs the bias–variance trade-off and determines whether a model underfits or overfits.In practice, selecting an appropriate regularisation strength is essential for building robust and interpretable classifiers. Logistic Regression, when properly regularised, remains a powerful and reliable tool for many classification tasks.

References

- Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.
- scikit-learn documentation: Logistic Regression.