# ML Lab 05

## Logistic Regression for Breast Cancer Prediction – Regularization

After uncommenting and rerunning the Logistic Regression model with varying values of C, the training and test accuracies remain the same (Training Accuracy: 0.9890, Test Accuracy: 0.9825). This suggests that the model was already trained with the best regularization parameter before, and rerunning it simply confirmed the same results.

**What Does C Represent in Logistic Regression?**

* C is the inverse of the regularization strength (C = 1/λ).
* Smaller C - Stronger regularization (more penalty on large coefficients).
* Larger C - Weaker regularization (less penalty, more flexibility in fitting data).

**Behavior of Training and Test Accuracy as C Varies**

1. For very small C (left side, strong regularization)

* Both training and test accuracy are low.
* The model is underfitting because strong regularization prevents it from learning complex patterns.

1. As C increases (moderate regularization)

* Accuracy improves as the model finds a good balance between bias and variance.
* The test accuracy peaks at around C ≈ 0.1 to 1, indicating the best generalization.

1. For very large C (right side, weak/no regularization)

* Training accuracy increases to nearly 100%, meaning the model fits the training data very well.
* Test accuracy drops significantly, showing that the model is overfitting (it memorizes training data but performs poorly on unseen data).

**Optimum Value for C (Best Regularization Strength)**

* From the plot, the optimal C value is around 0.1 to 1.
* This range gives the highest test accuracy (~0.9825) while keeping training accuracy high but not excessively overfitting.
* Choosing a very high C (e.g., 100 or 1000) causes severe overfitting, while choosing a very low C (e.g., 0.001) results in underfitting.

**Justification for Optimal C**

I aim to maximize test accuracy while keeping training accuracy reasonably high. The peak in test accuracy (around C = 0.1 to 1) shows the best trade-off between underfitting and overfitting. Beyond this, increasing C further hurts generalization.

