

CSAC 224 - MACHINE LEARNING

The Dataset

The classification dataset, entitled "Cancer Data," is from the Kaggle Datasets. It contains 570 instances and 30 attributes that determine whether a certain cancer is malignant (M) or benign (B). We removed several features except the radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and diagnosis, which are the only focus of my dataset exploration.

Data	columns (total 10 co.	lumns):	
#	Column	Non-Null Count	Dtype
Θ	radius_mean	569 non-null	float64
1	texture_mean	569 non-null	float64
2	perimeter_mean	569 non-null	float64
3	area_mean	569 non-null	float64
4	smoothness_mean	569 non-null	float64
5	compactness_mean	569 non-null	float64
6	concavity_mean	569 non-null	float64
7	concave points_mean	569 non-null	float64
8	symmetry_mean	569 non-null	float64
9	diagnosis	569 non-null	int64

Figure 1. Cancer Dataset Information

Logistic Regression Result

The diagnosis column is mapped with values of 0 for Benign and 1 for Malignant. The random noise is added for imperfection simulation of data entries as it should have behave in the real-world situation. The dataset is divided 80% for training set and 20% for the testing set. Now, the logistic regression model was trained with L1 (lasso) and L2 (ridge) regularizations to penalize our model. In the figure below are the classification reports for both penalties.

Classification	Report for precision	5	rization: f1-score	support	Classification	Report for precision	3	rization: f1-score	support
0 1	0.93 0.93	0.96 0.88	0.94 0.90	71 43	0 1	0.92 0.93	0.96 0.86	0.94 0.89	71 43
accuracy macro avg weighted avg	0.93 0.93	0.92 0.93	0.93 0.92 0.93	114 114 114	accuracy macro avg weighted avg	0.92 0.92	0.91 0.92	0.92 0.91 0.92	114 114 114

Figure 2-3. L1 and L2 Class Report

Analyzing the values , we say there is a tight difference between these two penalties implemented in training our model for the cancer dataset that we have (Fig. 1 and 2). By computing the accuracies of these models, our evaluation is 92.98% accuracy for L1 and 92.11% for L2. But if we do not use random noise for our dataset, the evaluation is 94.74% for L1 while 93.86% for L2 (this is for comparison only and not our final answer).

Moreover, the AUC value of the ROC Curve for L1 and L2 both sit comfortably at 0.98. This means that a higher AUC value (closer to 1.00) is a better indication of the logistic regression model performance.

1

¹ Taha, E. (2023). 'Cancer Data: Benign and Malignant Cancer Data'. https://www.kaggle.com/datasets/erdemtaha/cancer-data/data