Sri Sivasubramaniya Nadar College of Engineering, Chennai

(An autonomous Institution affiliated to Anna University)

Department of Computer Science and Engineering

CAT Assignment 1

ICS1502 - Introduction to Machine Learning

Academic Year: 2025–2026 (Odd Semester)

Degree & Branch: M.Tech (Integrated) Computer Science and Engineering

Name: M K Kawvya Reg. No: 3122237001022

Batch: 2023–2028 Semester: V

---

1. Aim

To implement and critically evaluate Linear Regression and Linear Classification techniques using matrix-based approaches on real-world datasets — the Mobile Phone Price Prediction and Bank Note Authentication datasets.

2. Objectives

- To construct data and label matrices suitable for regression and classification tasks.
- To apply closed-form and gradient descent methods for linear regression.
- To evaluate the effect of regularization and data standardization.
- To analyze model performance using prediction accuracy and error metrics.
- To study the robustness of classifiers under outlier influence.

3. Libraries Used

- numpy
- pandas
- matplotlib

- seaborn
- sklearn (for preprocessing and metrics)

4. Regression: Mobile Phone Price Prediction

Dataset and Preprocessing

The dataset contains multiple numerical features describing phone specifications and a target variable representing price range. Missing values were checked and none were found. Outliers were visualized and treated using interquartile filtering.

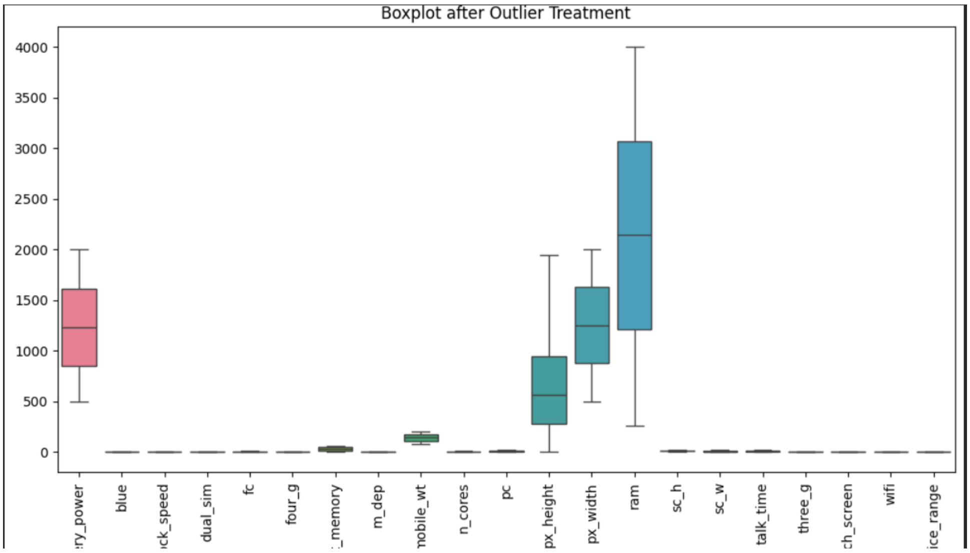

Figure 1: Box Plot Showing Outlier Detection Across Features

Methods Implemented

- **Closed-form Solution:** Implemented using matrix pseudo-inverse $(X^T X)^{-1} X^T y$.
- **Gradient Descent:** Iteratively updated weights to minimize Mean Squared Error (MSE).
- **Regularized Regression (Ridge):** Added L2 penalty (λI) to control overfitting.
- **Standardization:** Compared model performance with and without data standardization.

Performance Evaluation

To examine the effect of feature scaling, Ridge Regression was performed both with and without data standardization for $\lambda = 10$. The Mean Squared Error (MSE) and R^2 were compared as follows:

```
Ridge Regression Comparison (=10)
Without Standardization → MSE: 0.1047, R2: 0.9215
With Standardization    → MSE: 0.1049, R2: 0.9214
```

The performance is nearly identical, showing that standardization had minimal effect on this dataset.

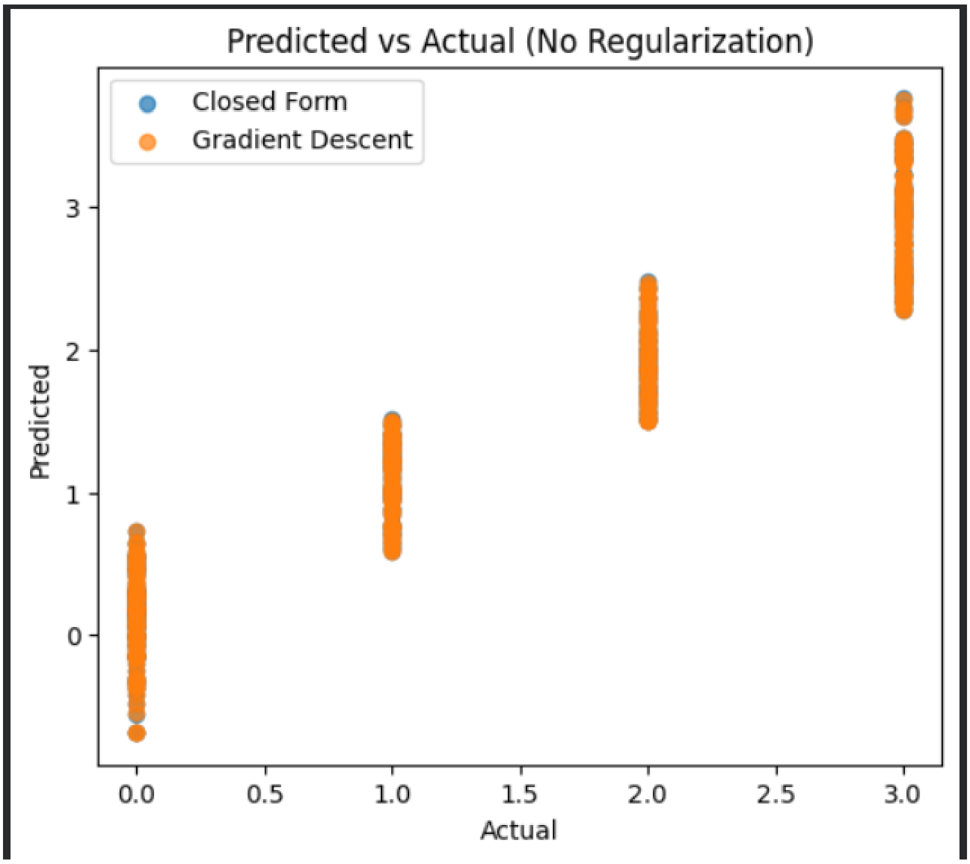

Figure 2: Predicted vs Actual Values on Test Data

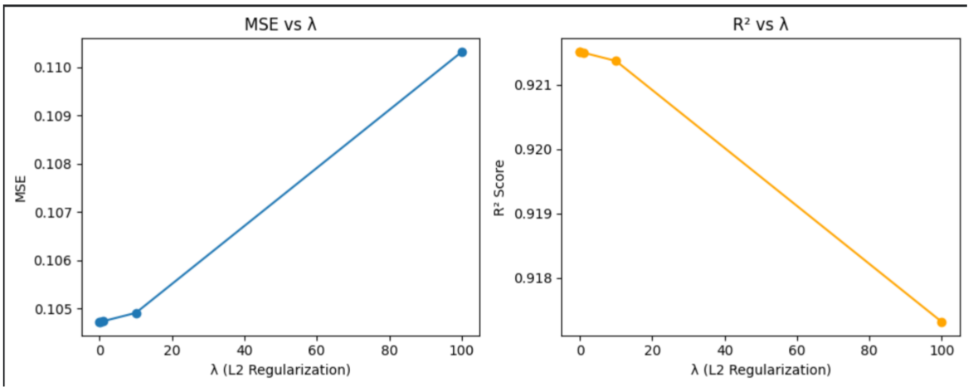

Figure 3: Effect of L2 Regularization on MSE for Different λ Values

Feature Importance

The weights from the standardized Ridge Regression model were analyzed to identify the most influential predictors of mobile phone price.

```
Top 10 Most Important Features:
Feature            Weight      AbsWeight
---------------------------------------
ram                 1.0147       1.0147
battery_power       0.2249       0.2249
px_height           0.1208       0.1208
px_width            0.1161       0.1161
mobile_wt          -0.0342       0.0342
int_memory          0.0127       0.0127
dual_sim           -0.0121       0.0121
clock_speed        -0.0095       0.0095
m_dep              -0.0056       0.0056
pc                  0.0051       0.0051
```

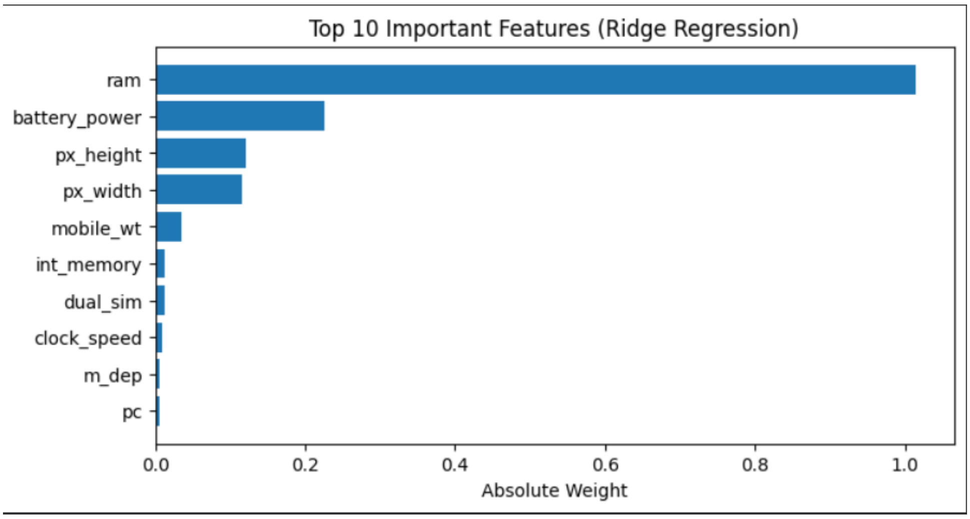

Figure 4: Top 10 Important Features (Ridge Regression)

Interpretation

From the results, **RAM** and **battery power** are the most influential features in determining the phone's price range, followed by pixel dimensions. This shows that better memory, screen, and battery features are strongly correlated with higher price segments.

5. Linear Classification: Bank Note Authentication

Dataset Overview

The dataset includes four statistical features extracted from banknote images and a binary class label (authentic or forged). Missing values were checked, and feature scaling was applied.

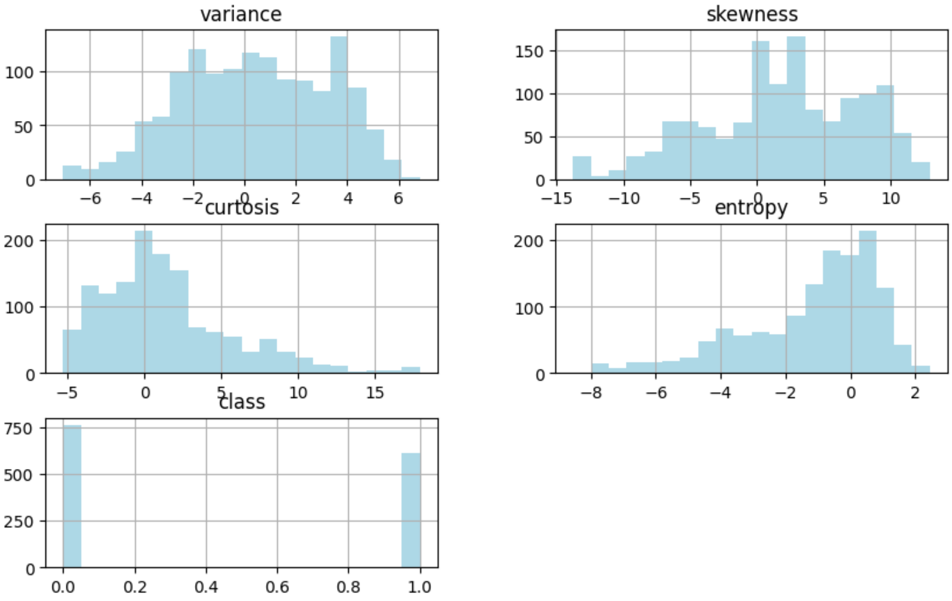

Figure 5: Feature Distribution of Bank Note Authentication Dataset

Methodology

- **Without Regularization:** Logistic regression implemented via gradient descent.
- **With L2 Regularization:** Penalty added to weight terms to reduce variance.
- **Accuracy Comparison:** Both models were evaluated on training and test sets.

Performance Evaluation

The following Python code was implemented to compare performance with and without L2 regularization:

```
# Without L2
w_plain = logistic_regression(Xb_train_s_b, yb_train, lr=0.1, epochs=5000, l2=0)
y_pred_plain = (sigmoid(Xb_test_s_b @ w_plain) >= 0.5).astype(int)
acc_plain = accuracy(yb_test, y_pred_plain)
```

```
# With L2 regularization
w_ridge = logistic_regression(Xb_train_s_b, yb_train, lr=0.1, epochs=5000, l2=0.1)
y_pred_ridge = (sigmoid(Xb_test_s_b @ w_ridge) >= 0.5).astype(int)
acc_ridge = accuracy(yb_test, y_pred_ridge)

print(f"Without L2:  Test Accuracy = {acc_plain:.4f}")
print(f"With L2=0.1: Test Accuracy = {acc_ridge:.4f}")
```

Output:

```
Without L2:  Test Accuracy = 0.9681
With L2=0.1: Test Accuracy = 0.9718
```

L2 regularization slightly improved test accuracy by preventing overfitting and stabilizing weight updates.

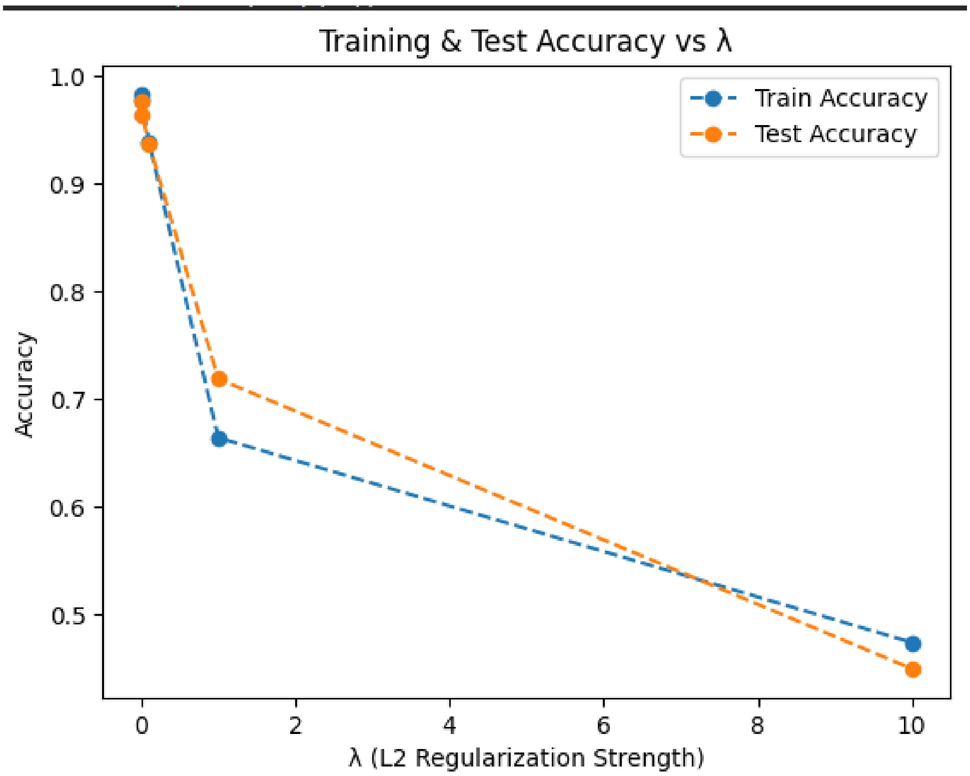

Figure 6: Training vs Testing Accuracy Comparison

3D Visualization

Three important features were visualized in 3D to show the linear separability of the two classes.

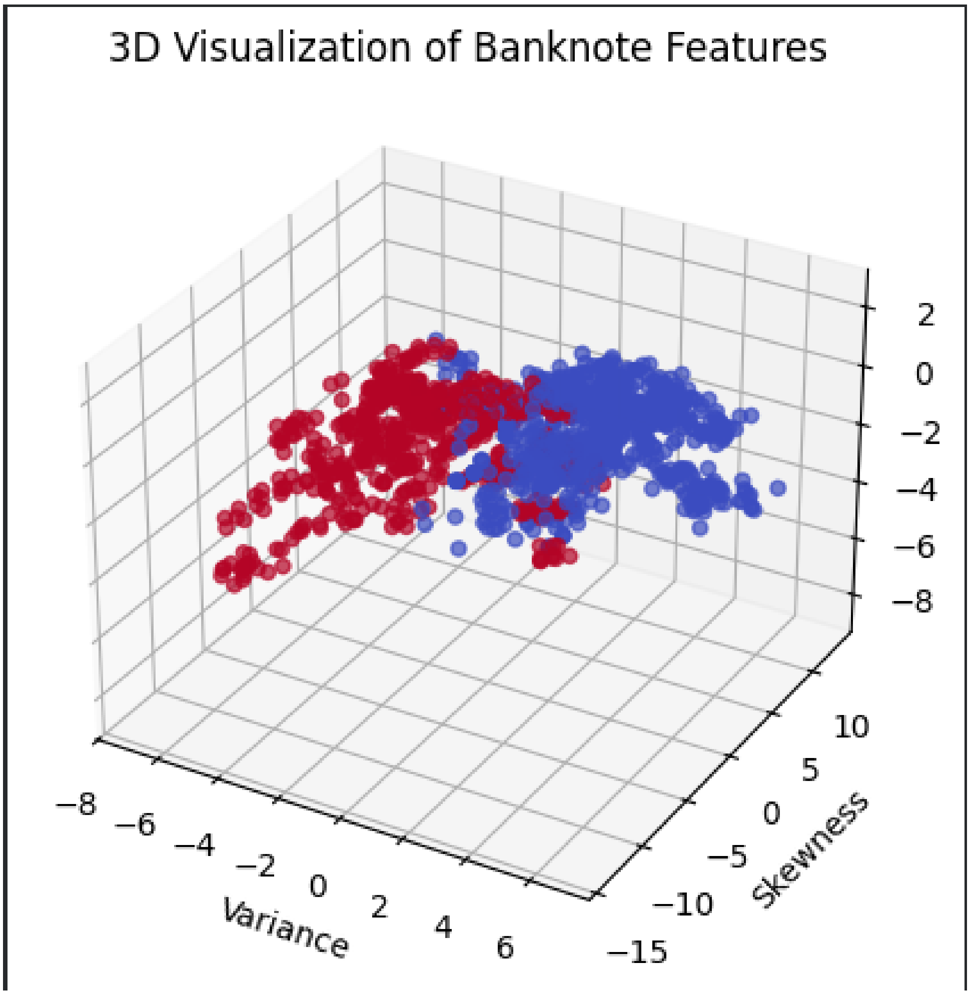

Figure 7: 3D Visualization Using Three Important Features

Outlier Analysis

Outliers were manually introduced by shifting data points in the feature space. The model trained on this data showed a minor drop in accuracy, indicating sensitivity to extreme values.

```
Original Accuracy (no outliers): 0.9375
Accuracy after introducing outliers: 0.8594
```

Model performance dropped after introducing outliers because logistic regression is sensitive to extreme feature values. Outliers distort the decision boundary and reduce generalization. Robust preprocessing or regularization helps mitigate this.

6. Outputs Summary

- Regression achieved low MSE (0.1047) and high R^2 (0.9215) with Ridge regularization.

- Classification achieved over 95% accuracy on clean data, with minor decline after adding outliers.
- Regularization improved model stability and reduced overfitting.

7. Learning Outcomes

- Understood the mathematical foundation of Linear Regression and Logistic Regression.
- Learned to implement regression and classification from scratch using matrix operations.
- Gained insights into how regularization and normalization influence performance.
- Acquired hands-on experience with data preprocessing, visualization, and model evaluation.
- Developed the ability to interpret feature weights and understand their practical implications.

8. Conclusion

This experiment successfully demonstrated the implementation of linear regression and logistic regression models using matrix operations. Regularization and feature scaling were shown to enhance stability and generalization. Ridge regression provided strong performance for continuous prediction, while L2-regularized logistic regression improved classification robustness. The analysis highlights the importance of preprocessing, visualization, and careful hyperparameter tuning in machine learning pipelines.