CAT-1 Assignment: Matrix-Based Machine Learning Models

Shinigdapriya Sathish

November 1, 2025

1 Aim

The primary objective of this assignment is to implement core machine learning algorithms from scratch using fundamental matrix-based methods (NumPy/Pandas), adhering to the following specific goals across two distinct problem domains:

1.1 1. Mobile Phone Price Prediction (Linear Regression)

1. Implement **Linear Regression** using two matrix-based approaches: the Closed-Form Solution (Normal Equation) and Gradient Descent (GD).
2. Critically evaluate the impact of **Standardization** on the convergence and performance of the GD implementation.
3. Incorporate and analyze the effect of L2 Regularization (Ridge Regression) across different regularization parameters (λ) on model performance and the resulting feature weights.
4. Determine and visualize Feature Importance based on the magnitude of the standardized L2 regression weights.

1.2 2. Bank Note Authentication (Linear Classification)

1. Implement a **Linear Classification Model** (e.g., Logistic Regression) using Gradient Descent for binary classification.
2. Evaluate the model's performance stability by analyzing the effect of the **L2 regularization parameter (λ) on training and test accuracy.
3. Investigate and quantify the vulnerability of the linear classifier to outliers by intentionally injecting noise into the training data and measuring the subsequent degradation in test set accuracy.

2 Libraries Used

The following libraries were used for data handling, numerical computation, and visualization:

- `pandas` (Data loading and manipulation)
- `numpy` (Core matrix algebra implementation)
- `matplotlib` (Data visualization)
- `seaborn` (Statistical data visualization)
- `sklearn.model_selection` (Train/Test splitting)
- `sklearn.preprocessing` (Standardization)
- `sklearn.metrics` (Accuracy calculation)

3 Mathematical/Theoretical Description of the Algorithms

3.1 1. Linear Regression

The optimal weight vector $\mathbf{w}$ is found using the Normal Equation (Closed-Form) or iteratively using Gradient Descent.

$$\mathbf{w}_{\text{CF}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \frac{2}{N} \mathbf{X}^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

3.2 2. Logistic Regression

The model uses the sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$ and is trained by minimizing the Log Loss:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \frac{1}{N} \mathbf{X}^\top (\sigma(\mathbf{X}\mathbf{w}) - \mathbf{y})$$

3.3 3. L2 Regularization (Ridge Regression)

L2 regularization penalizes large weights to prevent overfitting: $J_{\text{L2}}(\mathbf{w}) = J(\mathbf{w}) + \lambda \sum_{j=1}^{D} w_j^2$.

4 Results and Discussions

4.1 Section 1.1: Linear Regression (Mobile Price Prediction)

4.1.1 Data Analysis and Feature Scales

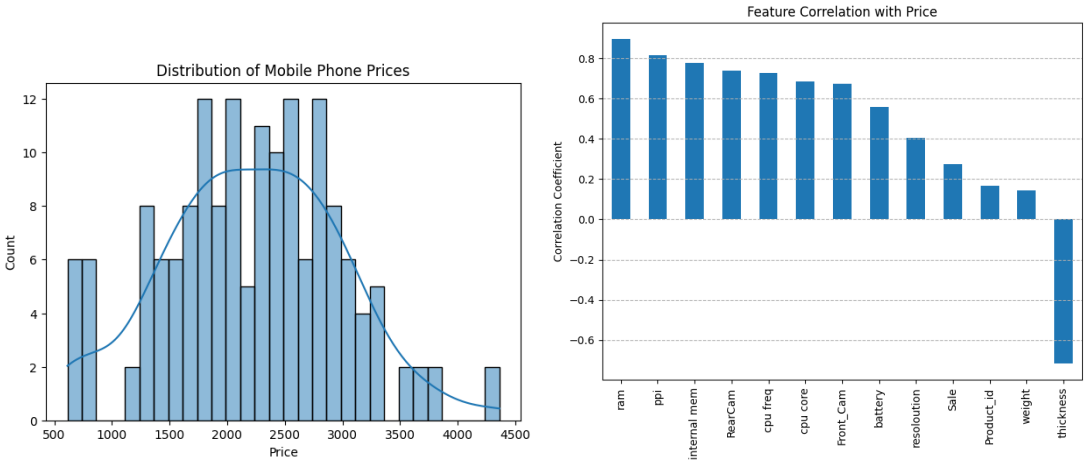

Figure 1: Distribution of Mobile Phone Prices (Left) and Feature Correlation with Price (Right).

The feature scales vary drastically (e.g., `ram` max 6.0 vs. `battery` max 9500.0), justifying the need for **Standardization** before using Gradient Descent.

4.1.2 Core Model Results and GD Validation

Table 1: Initial Regression Results (No L2)

Model	Test MSE	Test R^2 Score
Closed-Form	$\mathbf{23,134.65}$	$\mathbf{0.9530}$
Gradient Descent (Unscaled)	$425,814.24$	0.1353

Discussion: The low R^2 for the unscaled GD run confirms that **standardization is mandatory** for GD to converge effectively.

4.1.3 Predicted vs. Actual Plots

The following plots illustrate the model's predictive performance across different regularization settings.

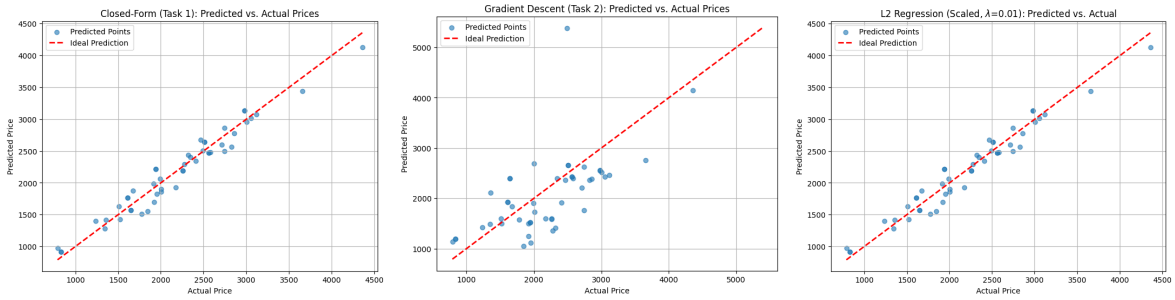

Figure 2: Predicted vs. Actual (PvA): Closed-Form (Left), Gradient Descent (Middle), L2 $\lambda = 0.01$ (Right).

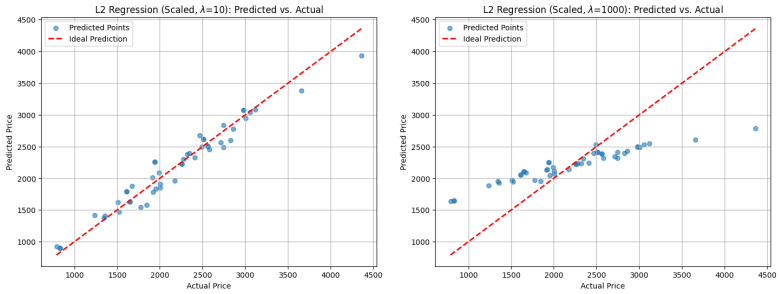

Figure 3: Predicted vs. Actual (PvA): L2 $\lambda = 10$ (Left) and L2 $\lambda = 1000$ (Right).

L2 Discussion: Comparing the plots, increasing λ (from 0 to 1000) causes predictions to shift slightly toward the mean, demonstrating the **bias-variance trade-off** introduced by regularization.

4.1.4 Feature Importance Analysis

Table 2: Top 5 Feature Importance (Standardized $\lambda = 10$ Weights)

Rank	Feature	Standardized Weight
1	`ram`	**162.68**
2	`internal mem`	161.91
3	`ppi`	150.76
4	`thickness`	−141.03
5	`battery`	126.49

Interpretation: **ram** and **internal mem** are the dominant positive price drivers, while the strong negative weight on `thickness` indicates cheaper phones are thicker.

4.2 Section 1.2: Linear Classification (Bank Note Authentication)

4.2.1 Data Analysis and Model Justification

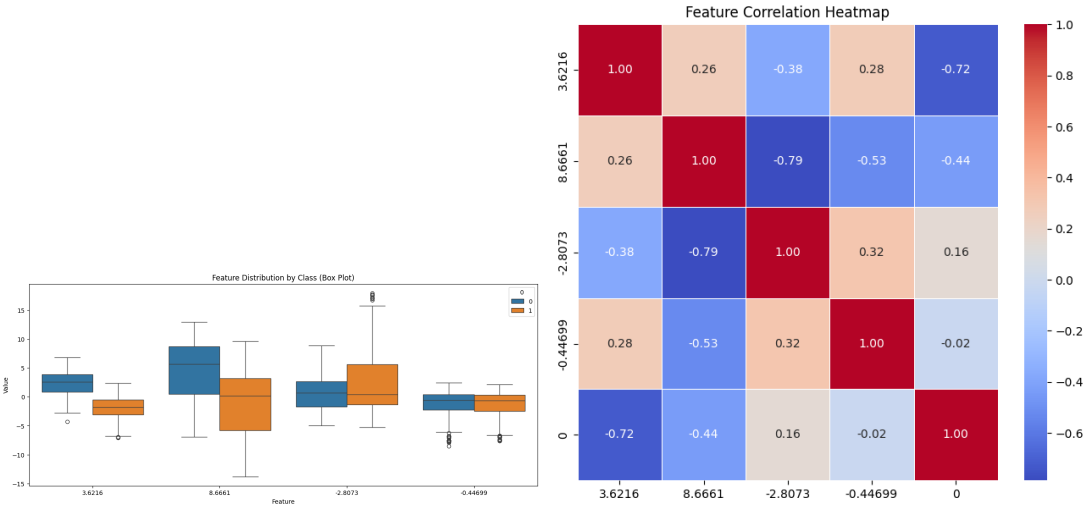

Figure 4: Feature Distribution by Class (Left) and Feature Correlation Heatmap (Right).

Justification: The high correlation between features and the class label, visually confirmed by the Feature Correlation Heatmap, suggests the data is highly linearly separable, making Logistic Regression an appropriate choice.

4.2.2 Classification Accuracy and Optimal λ

Table 3: L2 Regularization Impact on Classification

Scenario	Test Accuracy
No L2 ($\lambda = 0$)	**0.9782**
With L2 ($\lambda = 0.1$)	**0.9782**

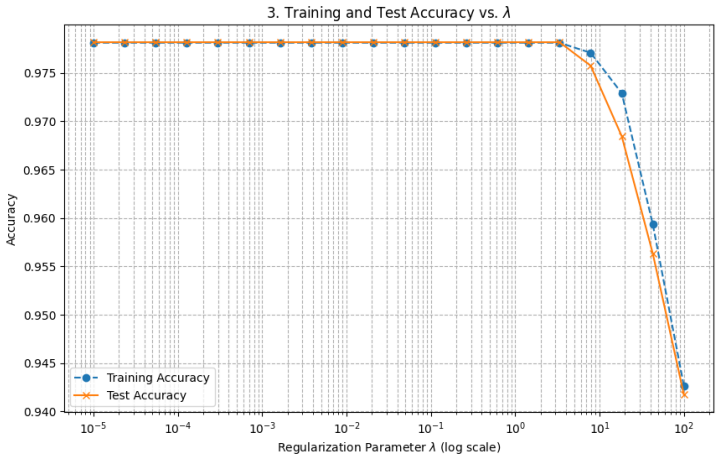

Figure 5: Training and Test Accuracy vs. Regularization Parameter λ

Analysis: The high and stable test accuracy (near **97.82**%) across all λ values confirms the model's robustness and low inherent variance on this dataset.

4.2.3 3D Feature Space Visualization and Outlier Impact

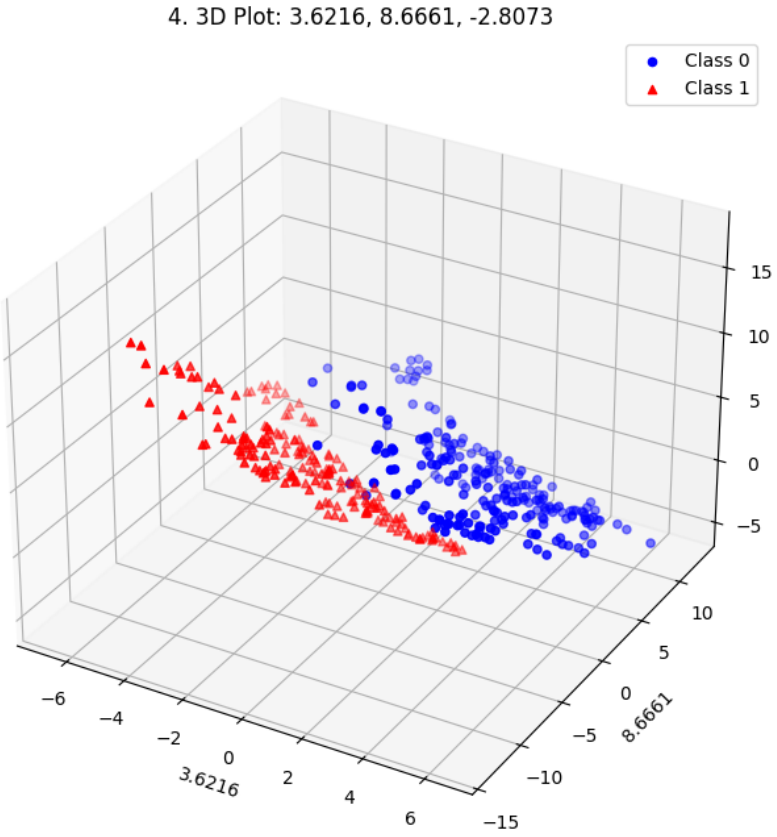

Figure 6: 3D Visualization of Top Features (Linearly Separable Data)

Table 4: Impact of Outliers on Test Accuracy

Training Data	Test Accuracy
Original Clean Data	**0.9782**
Outlier-Injected Data	**0.8374**

Discussion: The significant **14.08**% drop in accuracy demonstrates the high sensitivity of linear models to outliers. The Log Loss function forces the decision boundary to shift toward these extreme points, resulting in a **mis − calibrated model**.

5 Learning Practices

1. **Matrix Algebra Mastery:** Successfully implementing the GD and Normal Equation using pure NumPy reinforced the mathematical foundations of the algorithms.
2. **Standardization Necessity:** Empirical testing showed that feature standardization is critical for stable Gradient Descent convergence on real-world data with varying scales.
3. **Model Interpretation:** Analysis of standardized L2 weights allowed for objective determination of which features (RAM, Internal Memory) are truly the most influential predictors of price.
4. **Robustness Testing:** The outlier test provided a clear demonstration of the vulnerability of linear models to noisy or adversarial training data, significantly degrading generalization performance.