CAT Assignment 1 – Theory Report

ICS1502 – Introduction to Machine Learning

Sri Sivasubramaniya Nadar College of Engineering
Department of Computer Science and Engineering

Name: Pranavah Varun M V
Roll Number: 3122237001039

Academic Year: 2025–2026 (Odd)
Batch: 2023–2028

Note on Submission

All required plots, outputs, and implementation details are included in the accompanying Jupyter Notebook. This report focuses only on the theoretical analysis, explanations, and conclusions as required.

1. Regression – Mobile Phone Price Prediction

We evaluated **Linear Regression** for predicting mobile phone prices using both the closed-form solution (normal equation) and gradient descent.

Data Representation

The dataset was divided into training and test sets to ensure fair evaluation. The *design matrix* X $(m \times n)$ contained feature values of mobile phones, while the label vector y contained price values. The parameter vector θ $(n \times 1)$ represented weights. Each parameter indicated how strongly its corresponding feature influenced the prediction.

Closed-Form vs Gradient Descent

The closed-form solution directly solved $\theta = (X^T X)^{-1} X^T y$. Gradient descent optimized weights iteratively by minimizing mean squared error. Both methods produced consistent results, validating correctness.

With L2 Regularization

Ridge regression was applied by modifying the loss:

$$J(\theta) = \frac{1}{m}\|X\theta - y\|^2 + \lambda\|\theta\|^2$$

This penalized large weights, reducing overfitting and variance.

Effect of Standardization

Without feature scaling, attributes with large ranges dominated the model. Standardization improved ridge regression by giving all features equal weight, leading to more balanced predictions.

Error Analysis and Performance

We used Mean Squared Error (MSE) and R^2 to evaluate performance. Regularized models achieved slightly lower training error variance and were more stable under different λ values.

Feature Importance

From the magnitude of ridge regression weights, features such as *RAM* and *internal storage* had the greatest influence on predicted prices, while minor specifications had smaller effects.

Conclusion: Linear regression with appropriate preprocessing and regularization is effective for modeling mobile phone prices.

2. Linear Classification – Bank Note Authentication

We applied a linear classifier on the Bank Note Authentication dataset.

Train-Test Split

Data was divided into training and testing sets. This allowed evaluation of generalization on unseen examples.

L2 Regularization

We trained models with and without L2 penalty. With regularization, weights were smaller and the gap between train and test accuracy reduced. Training accuracy decreased slightly while test accuracy stayed nearly constant, showing good generalization.

Accuracy vs λ

Plots of training and test accuracy against λ (log scale) showed that higher λ reduced training accuracy while test accuracy remained stable. This indicates that the model already generalized well without much regularization.

Visualization

A 3D scatter plot using three features (variance, skewness, and curtosis) revealed clear separation between classes, supporting the suitability of a linear boundary.

Impact of Outliers

We artificially introduced outliers by shifting some data points. The classifier trained on this modified dataset showed reduced performance and less stable decision boundaries. This demonstrated the sensitivity of linear models to noisy data.

Conclusion: Linear classifiers performed very well on this dataset, achieving $\sim$ 98–99% accuracy. Regularization improved robustness, while visualization confirmed that linear decision boundaries were appropriate.

3. Summary

Both regression and classification experiments confirmed the effectiveness of linear models when combined with proper regularization. Regression highlighted the importance of preprocessing and feature scaling, while classification demonstrated the impact of regularization and outliers. Together, these tasks reinforced theoretical understanding and practical application of linear approaches in machine learning.