Sri Sivasubramaniya Nadar College of Engineering, Chennai

(An autonomous Institution affiliated to Anna University)

Department of Computer Science and Engineering

CAT Assignment 1

Degree & Branch	M. Tech (Integrated) Computer Science & Engineering	Semester	V
Subject Code & Name	ICS1502 & Introduction to Machine Learning		
Academic year	2025-2026 (Odd)	Batch:2023-2028	**Due date:28-08-2025**

Moogambigai A - 3122237001027

Theory Answers – Regression and Classification

1. Regression (Matrix Approach)

Data Representation

In regression, we represent our data in matrix form:

- **X** is the data matrix with shape $(m \times n)$, where m is the number of samples and n is the number of features.
- We add a column of ones to X to include the intercept (bias), giving us **X'** of shape $(m \times (n+1))$.
- **y** is the target (price) column with shape $(m \times 1)$.
- **θ (theta)** is the parameter or weight vector of shape $((n+1) \times 1)$.

The prediction is made as: $\hat{y} = X'\theta$

Closed-Form Solution (Normal Equation)

The normal equation gives a direct formula to find the best θ:
$\theta = (X'^T X' + \lambda I)^{-1} X'^T y$

- When $\lambda = 0$, it's the **ordinary least squares** (no regularization).
- When $\lambda > 0$, it's **ridge regression (L2 regularization)**, which helps reduce overfitting by keeping weights small.

Gradient Descent

Gradient descent is an **iterative method** that adjusts θ step by step to minimize the error.

The update rule is:
$\boldsymbol{\theta \leftarrow \theta - \alpha}$ *** gradient**,
where α is the learning rate.

For regression with L2 regularization, the gradient is:
$$\mathbf{(1/m)\, X'^T(X'\theta - y) + (\lambda/m)[0;\, \theta_1..\theta_n]}$$

It slowly learns the best parameters over many iterations.

Error Analysis and Performance

We evaluate regression models using:

- **MSE (Mean Squared Error):** Measures average squared difference between actual and predicted values.
- **RMSE (Root Mean Squared Error):** Easier to interpret, same scale as the target.
- **R^2 (Coefficient of Determination):** Measures how well the model explains the variance in the target.

We test the model on unseen data (test set) to check how well it generalizes.
Plotting **predicted vs. actual** values helps us see if predictions follow the true pattern or if there's bias.

Standardization and Regularization

- **Standardization** (scaling features to zero mean and unit variance) helps algorithms like gradient descent converge faster and ensures that regularization treats all features fairly.
- Without standardization, features with larger numeric ranges can dominate the model and make regularization less effective.

2. Classification (Bank Note Authentication)

Model Choice

For this dataset, **Logistic Regression** (a linear classification model) is a good fit because the classes can be separated fairly well using linear boundaries.

Regularization Effect

- **Without regularization**, the model may overfit to noise in the training data.
- **With L2 regularization**, the model becomes more stable and generalizes better by shrinking large weights.

Performance Evaluation

We compare **training** and **testing accuracy** to check for overfitting.
We can also plot **accuracy vs λ** (regularization strength) to see how the model performs for different penalty values.

Outliers

Outliers are extreme data points that don't follow the normal pattern.
When we intentionally add outliers:

- The model accuracy usually decreases.
- Logistic regression's decision boundary may shift incorrectly.

This shows that **outliers can strongly affect linear models**, so it's important to handle them before training.