Project Repository Report

Comparison Analysis of Regression Techniques Using Real World Datasets

Author: Revanth Krishna Sai Teegala

Introduction

This project repository focuses on the application and comparative analysis of various regression techniques using real-world datasets. The primary objective is to determine which regression technique performs best in terms of accuracy for specific datasets. The regression techniques analyzed include linear regression, multiple regression, polynomial regression, and logistic regression. The datasets used for this analysis are Iris Plants, Breast Cancer, Pima Indians Diabetes, Heart Disease, and Car Evaluation.

Description

Machine learning has significantly impacted numerous fields, such as agriculture, healthcare, and marketing, by offering solutions to real-life problems. This repository aims to provide valuable insights into different machine learning regression techniques and their application on real-world datasets.

Datasets

1. Iris Plants Dataset:
   - Features: Sepal length, Sepal width, Petal length, Petal width
   - Classes: Iris setosa, Iris versicolor, Iris virginica

2. Breast Cancer Dataset:
   - Features: Various numerical attributes
   - Class: Presence of breast cancer (binary classification)

3. Pima Indians Diabetes Dataset:
   - Features: Various medical attributes
   - Class: Diabetes condition (binary classification)

4. Heart Disease Dataset:
   - Features: Medical attributes related to heart disease
   - Class: Risk of heart attack (binary classification)

5. Car Evaluation Dataset:
   - Features: Buying price, Maintenance price, Number of doors, Capacity, Luggage boot size, Safety, Acceptability
   - Class: Car acceptability

Models

1. Linear Regression: Used to predict continuous numerical values.
2. Multiple Regression: An extension of linear regression that uses multiple predictors.
3. Polynomial Regression: Used when the relationship between variables is non-linear.
4. Logistic Regression: Used for binary classification tasks.

Methodology

1. Data Preprocessing:
   - Converting categorical values to numerical values.
   - Handling missing values.
   - Splitting data into training and testing sets.

2. Model Training and Evaluation:
   - Training models on the training set.
   - Making predictions on the test set.
   - Evaluating model performance using accuracy, confusion matrix, mean squared error, etc.

Results

The analysis revealed that no single regression model is universally superior. However, logistic regression generally showed high accuracy across various datasets, especially for classification tasks. Multiple regression performed well on datasets with multiple predictors. Polynomial regression was effective for datasets with complex patterns, and linear regression served as a good baseline for comparison.

Key Findings

- Iris Plants: Highest accuracy with Logistic Regression (0.967).
- Breast Cancer: Highest accuracy with Multiple Regression (0.973).
- Pima Indians Diabetes: Highest accuracy with Logistic Regression (0.77).
- Heart Disease: Highest accuracy with Polynomial Regression (0.916).
- Car Evaluation: Highest accuracy with Logistic Regression (0.7).

Conclusion

This project emphasizes the importance of selecting an appropriate regression model based on the dataset's characteristics. Logistic regression generally performs well for classification tasks, while multiple regression is effective for datasets with multiple predictors. Polynomial regression is suitable for datasets with non-linear relationships. Linear regression serves as a simple yet effective baseline model.

The findings from this project can guide the selection of regression techniques for various real-world applications, ensuring optimal model performance.

Repository Contents

- Code: Implementation of regression models.
- Datasets: Real-world datasets used for analysis.
- Documentation: Detailed explanation of methods and results.
- Reports: Comprehensive analysis and comparison of regression techniques.