Development of a Machine Learning Library from Scratch

Sukrat Singh
IIT (ISM) Dhanbad

Description

This project aims to develop a machine learning library by implementing foundational algorithms from scratch, providing a deep understanding of the mathematical and computational underpinnings. The library will support algorithms like **Linear Regression, Logistic Regression, K-Nearest Neighbors, K-Means Clustering, Decision Trees, and N-Layer Neural Networks**. The goal is to train these algorithms on provided datasets, evaluate their performance on unseen data, and document the process thoroughly.

Features

- **Linear Regression and Polynomial Regression**: Implementation of models to capture both linear and non-linear trends.
- **Logistic Regression**: Binary classification using the sigmoid function and cross-entropy loss.
- **K-Nearest Neighbors (KNN)**: Distance-based classification algorithm with flexible metrics.
- **K-Means Clustering**: Unsupervised learning algorithm for data segmentation.
- **Decision Trees**: Tree-based learning with support for classification and regression.
- **N-Layer Neural Network**: Fully connected neural networks with customizable layers and activation functions.

Technology Stack

- **NumPy**: For matrix operations and numerical computations.
- **Pandas**: For data manipulation and preprocessing.
- **Matplotlib**: For visualization of data and training progress.
- **Python**: Primary programming language.
- **Jupyter Notebook/Google Colab**: For development and demonstration.

Brief Implementation Details

The project will involve the following steps:

1. **Mathematical Understanding**: Analyze and understand the theoretical basis of each algorithm.
2. **Code Implementation**: Develop each algorithm using NumPy for numerical computations.
3. **Validation**: Train and evaluate models on the given dataset, ensuring robust performance metrics.
4. **Visualization**: Plot training logs, decision boundaries, and performance metrics to gain insights.
5. **Documentation**: Prepare a detailed report with training logs, hyperparameters, and experiment results.

Timeline

Phase	**Tasks**	**Timeline**
Phase 1	Linear Regression (with Polynomial Regression) and Logistic Regression	7 Dec - 13 Dec
Phase 2	K-Nearest Neighbors (KNN) and K-Means Clustering	14 Dec - 20 Dec
Phase 3	Decision Trees	21 Dec - 26 Dec
Phase 4	N-Layer Neural Network	27 Dec - 31 Dec
Phase 5	Fine-tuning, integration of modules, and testing on datasets	1 Jan - 3 Jan
Phase 6	Report preparation: Documenting methodologies, experiments, visualizations, and final results	4 Jan - 5 Jan

About Me

- **Personal Details**: Sukrat Singh, 24JE0702, Pursuing Bachelor of Technology in Computer Science and Engineering.
- **Why Should I Be Selected?**: I have a keen interest in machine learning and a good foundation in programming and mathematics. I am motivated to learn the nuances of ML by implementing foundational algorithms from scratch.
- **Commitments During the Program**: I can dedicate an average of 4-5 hours daily to this project. I will communicate any planned absences or reduced availability in advance.