Linear Regression with Regularization Homework

This zipfile contains Python code for performing linear regression with regularization (L2 regularization) on multiple datasets of varying sizes. The code calculates and visualizes the Mean Squared Error (MSE) for different values of lambda (λ) ranging from 0 to 150. It also contains the written answer to question 1 in one pdf, and the typed answers to 2,3 in another pdf.

Datasets

The code uses the following datasets:

Dataset 1: Train-100-10 (100 rows, 10 columns)
Dataset 2: Train-100-100 (100 rows, 100 columns)
Dataset 3: Train-1000-100 (1000 rows, 100 columns)

For Dataset 3, we also split it into smaller subsets:

- **Dataset 4:** Train-50(1000)-100 (50 rows from Dataset 3)
- **Dataset 5:** Train-100(1000)-100 (100 rows from Dataset 3)
- **Dataset 6:** Train-150(1000)-100 (150 rows from Dataset 3)

Code Structure

The code is organized into the following sections:

Data Loading and Preprocessing: Load and preprocess the datasets, adding an intercept term, and splitting the data when necessary.

Weight Calculation: Calculate the weights of linear regression with varying lambda values, ranging from 0 to 150.

Mean Squared Error (MSE) Calculation: Calculate the MSE for both the training and test datasets for each lambda value.

Cross-Validation (CV): Perform k-fold cross-validation (k=10) to find the optimal lambda value that minimizes the test MSE for each dataset.

Learning Curve: Generate learning curves to visualize the effect of training set size on MSE for different lambda values.

Results and Visualization

The code produces visualizations of the MSE values for different datasets, lambda values, and learning curves. It also identifies the lambda value that results in the lowest test MSE for each dataset.

How to Run

- Ensure you have Python, NumPy, Matplotlib, and Pandas installed on your system.
- Place your dataset files (e.g., train-100-10.csv, test-100-10.csv, etc.) in the same directory as this code.
- Run the code to perform linear regression with regularization and generate the visualizations.

Function Documentation

1. **dataframe_to_matrix(df):** Converts a Pandas DataFrame to matrices for features (X) and targets (Y).
2. **convert_dataset_to_matrices(train_df, test_df, split_df):** Converts train, test, and split DataFrames to matrices for X and Y.
3. **weight_calculation(X_train, Y_train, lambda_end, lambda_start=0):** Calculates the weights of linear regression with varying lambda values.
4. **mean_squared_error(X_train, weights, Y_train):** Calculates the Mean Squared Error (MSE) for linear regression predictions.

5. **learning_curve(x_matrix_train, y_matrix_train, x_matrix_test, y_matrix_test, rep, size):** Generates learning curves to visualize MSE with varying training set sizes and lambda values.