Assignment Clarification

Aakash Kumar

October 2024

Clarification on Programming Assignment 2

1. Creation of Training Data and Test Data

Response: Ignore the "FIXME" note. The training and test data handling will be specified elsewhere.

2. P1: RMSE Function

Response: Implement the RMSE function using both methods (Python `sklearn`) and then run experiments by defining values for y and $\hat{y}$. You should get identical results from both functions.

3. P2: Implement Linear Predictor

Response: The linear predictor follows the formula $\hat{y} = \theta_1 \cdot x + \theta_0$. Define values of θ_0 and θ_1, and run the function with either your own input (`x`) or use `x_train` / `x_test`.

4. P3: Implement Grid Search

Response: The grid search will iterate over possible values of θ_0 and θ_1 using a nested loop. Use the `predict` function and the error function to find the combination of θ that yields the lowest error. Report:

- Lowest training error
- Best θ
- Training time
- Test error after applying the best θ on `x_test`

5. P4: Random Search

Response: Loop for around 1000 iterations, randomly selecting θ values each time, you define your own range and randomly select the theta values form that range for each iteration:

- Predict using the random θ on the training set
- Calculate the error

Report:

- Lowest training error
- Best θ
- Training time
- Test error using the best θ

6. P5: Using Sklearn Library

Response: Implement linear regression using the `LinearRegression` class from `sklearn`. Report:

- Training time
- Training error
- Best θ
- Test error

Performance Discussion

Example format for discussing results:

Note: you may observe different result and you can reach to a completely different conclusion and that is fine.

Grid Search

- Best θ: [10.0, 5.0]
- Training Error: 1,024,069.74
- Test Error: 1,002,601.11
- Training Time: 0.001 seconds
- Conclusion: Fast, but high error rates and less optimal θ.

Random Search

- Best θ: [5.90, 9.98]
- Training Error: 1,023,207.17
- Test Error: 1,001,768.79
- Training Time: 0.0488 seconds
- Conclusion: Better than grid search, but slower and less accurate than `sklearn`.

Sklearn

- Best θ: (8579.51, 4418.09)
- Training Error: 508,184.67
- Test Error: 503,348.04
- Training Time: 0.001 seconds
- Conclusion: Fastest and most accurate, outperforming both grid and random search.

Conclusion: The `sklearn` Linear Regression implementation outperformed both the grid search and random search approaches in terms of accuracy and training time. It is highly optimized for this type of problem, leveraging well-established numerical methods to minimize the error, and is thus recommended for practical applications over grid and random search for linear regression tasks.