Home Assignment – 5

Aarav Aryaman, 220012

Problem Statement:

System Identification refers to the technique of discovering governing equations of a system from experimental data. For scientists and engineers, system identification is a major application area of machine learning.

In this homework, we will find the governing equation of a simple pendulum from experimental data. The data are presented in the pendulum_data.csv file. The data file contains three columns: theta (angular displacement), theta_dot (angular_velocity), theta_double_dot (angular acceleration). A row of the data file indicates the angular displacement (θ), angular velocity (θ′) and angular acceleration(θ″) of the pendulum at a given time instant.

Our goal is to discover the governing equation of the pendulum. In general, θ″ could be a function of θ, sin(θ), θ′, θ'^2.

(a) Calculate correlation matrix, scatter plots and any other related metrics to qualitatively propose possible hypotheses and create the hypotheses space.

(b) Using linear or nonlinear regression with ridge regularization find appropriate parameters.

(c) Use cross validation to finalize your hypothesis.

Solution Procedure:

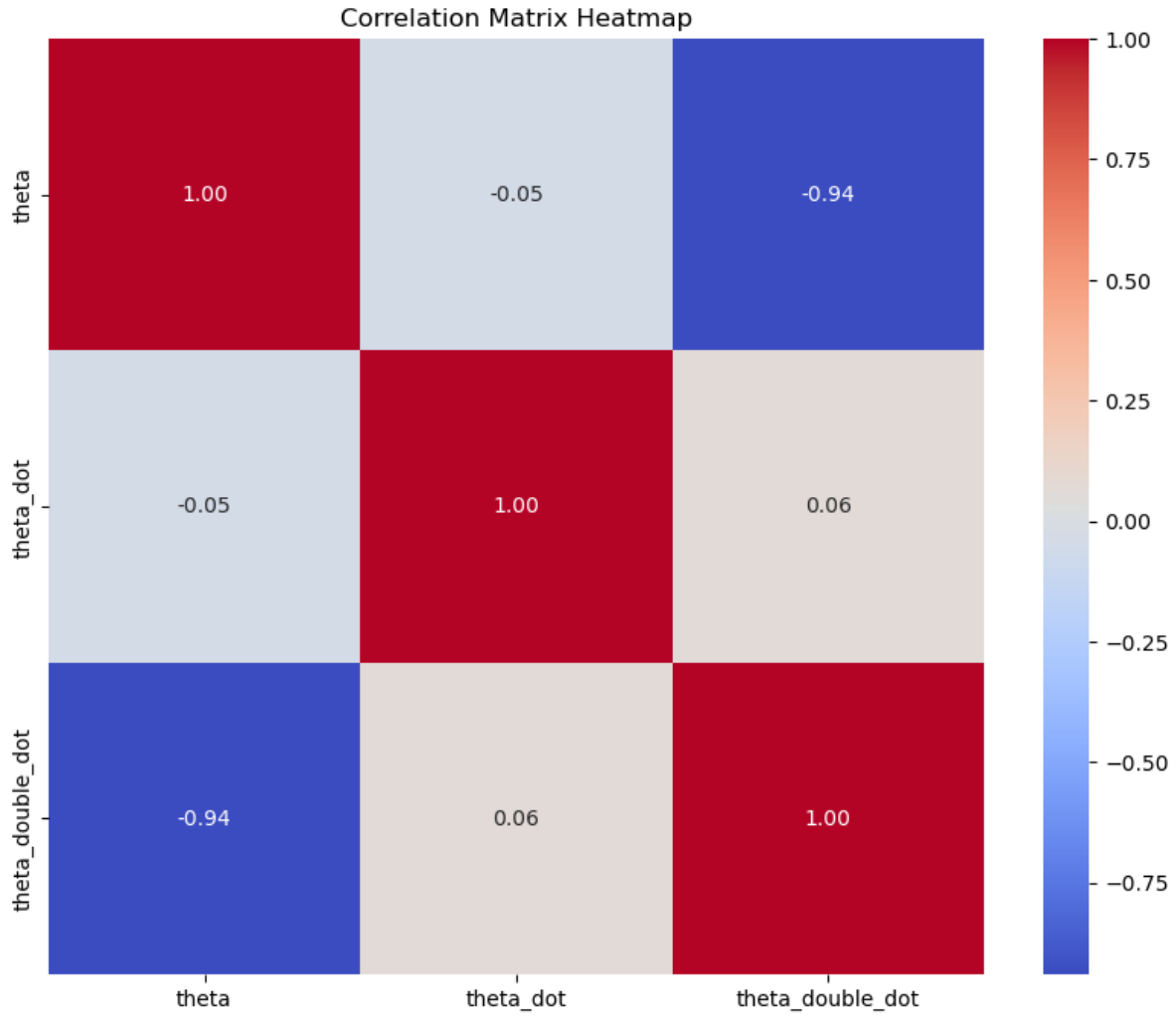
1. **Data Acquisition:**
    - The analysis begins with importing the given dataset (pendulum_data.csv).
2. **Data Analysis:**
    - The correlation matrix is computed to identify the relationships between the variables.
    - Scatter plots are generated to visually explore the relationships between angular acceleration (θ″) and potential predictor variables, including θ, sin(θ), θ′, θ'^2.
3. **Line Hypotheses:**
    - Three different line hypotheses are proposed:
        - **Line-1:** θ″ = a0 + a1 * θ + a2 * sin(θ) + a3 * θ′ + a4 * θ'^2 (Linear)
        - **Line-2:** θ″ = b0 + b1 * θ + b2 * sin(θ) + b3 * θ′ (Linear)
        - **Line-3:** θ″ = c0 + c1 * θ + c2 * (θ′ * sin(θ)) + c3 * θ'^2 (Non-linear)
4. **Model Training:**
    - Each hypothesis is trained using Ridge regression, which includes regularization to mitigate overfitting for both linear and non-linear cases. Prebuilt libraries are used for this purpose that are imported from sklearn.
5. **Cross-Validation:**

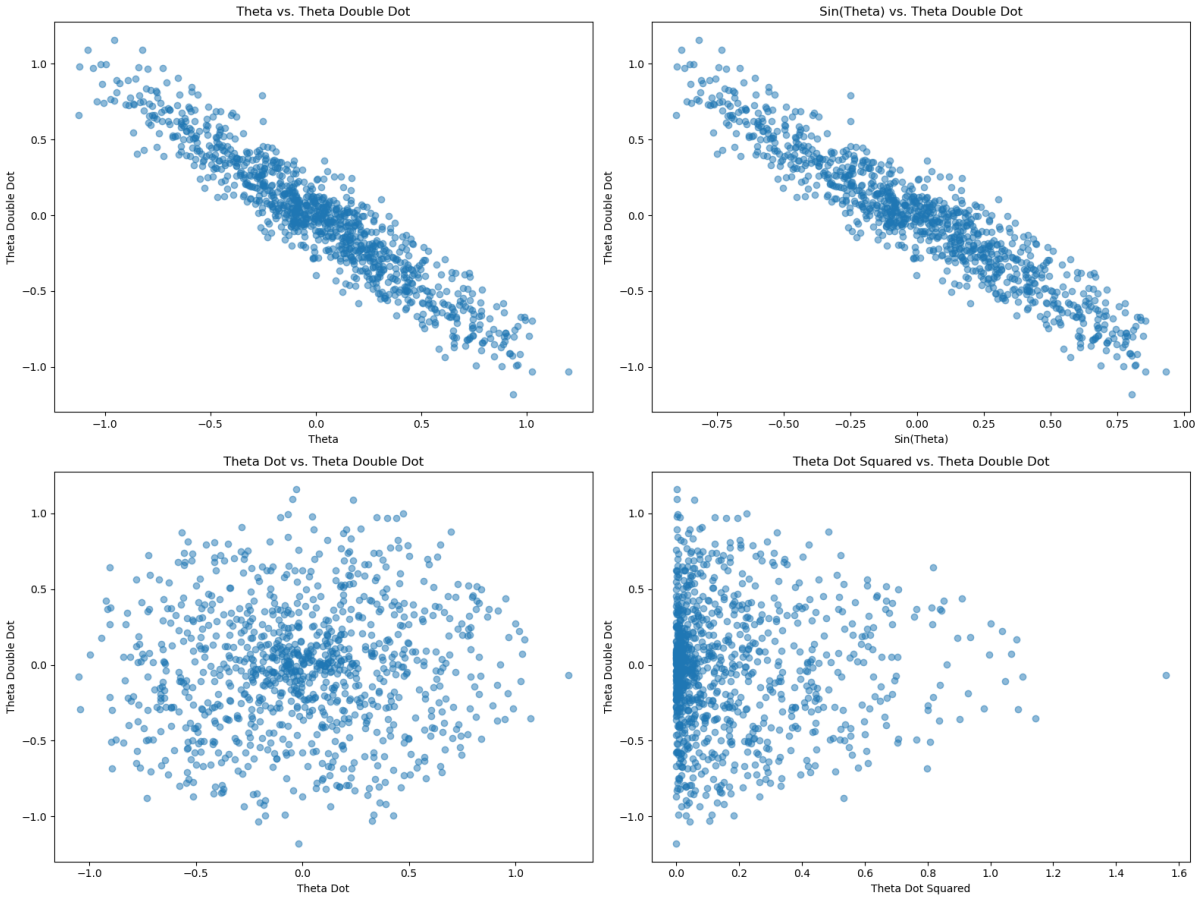
- A 5-fold cross-validation is performed to evaluate the models' performance, using metrics such as Mean Squared Error (MSE) and Coefficient of Determination (R^2). Prebuilt libraries are used for this purpose imported from sklearn.

6. **Result Analysis:**

   - The models are compared based on their average MSE (minimum) and R^2 scores (maximum) to identify the best-performing model. Here, **Line-1** is the best performing.

Plots:





Results and Discussion:

1. **Coefficients and Intercepts:**
    - **Line-1:**
        - Output the regression equation derived from the coefficients a1, a2, a3, a4 and intercept a0.
    - **Line-2:**
        - Output the regression equation derived from the coefficients b1, b2, b3 and intercept b0.
    - **Line-3:**
        - Output the regression equation derived from the coefficients c1, c2, c3 and intercept c0.
2. **Cross-Validation Results:**
    - Calculate the average MSE and R^2 values for each line hypothesis. From this, we find the best-performing line model as **Line-1**.
    - The performance metrics from cross-validation show that Line-1 achieved the best balance between complexity and predictive accuracy, indicating that the additional features provide significant explanatory power.

- Line-2 simplifies the relationship by omitting θ'^2. This model still maintains a solid performance, suggesting that the linear relationships between the variables can sufficiently explain a portion of the angular acceleration, but may overlook certain dynamics that become more pronounced at higher velocities.
- Line-3 introduces non-linear interaction between (θ' * sin(θ)). This model's performance was less favourable compared to both Line-1 and Line-2, suggesting that the inclusion of squared terms (θ'^2) may capture the underlying physics more effectively.

Conclusion:

In this analysis, we successfully modelled the dynamics of a pendulum using Ridge regression. By evaluating three distinct hypotheses, we were able to capture both linear and non-linear interactions, highlighting the significance of various features in predicting angular acceleration and determined the best model based on cross-validation metrics.

The results demonstrated that including a comprehensive set of features—particularly the square of angular velocity—greatly enhanced the model's performance. The first model, which incorporated both θ and θ'^2, yielded the best predictive accuracy. Additionally, the second model illustrated that simpler relationships can still provide substantial explanatory power, while the third model emphasized the value of non-linear interactions in complex dynamic systems.

Through the application of Ridge regression, we effectively mitigated the risk of overfitting while maintaining predictive power. This analysis not only provides insight into the physical behaviour of pendulum motion but also establishes a robust methodology for modelling similar dynamic systems.