Interactive Clustering and Regression Data Interface (ICARDI)

Ahmet Enes Topçu, 150210310
Alihan Esen, 150220338
Bedirhan Öztürk, 150210321

June 2024

Project Repository

The project repository can be found at this link.

Abstract

Our *Interactive Clustering and Regression Data Interface (ICARDI)* aims to provide an intuitive and interactive platform for users to perform regression and clustering on their datasets. The tool is designed to be user-friendly, requiring minimal coding or mathematical knowledge. It supports multiple methods for data input, including direct graph clicking, textbox entry, and CSV file import. The tool offers linear regression, polynomial regression, K-means clustering, and agglomerative clustering, with features to automatically determine optimal parameters such as polynomial degree and the number of clusters.

Problem Statement

The primary motivation for this project was to create a comprehensive, user-friendly tool for regression and clustering that can be easily used by individuals with little to no background in coding or mathematics. Existing tools often lack simplicity or do not combine both regression and clustering in a single platform. This tool aims to fill that gap by providing an all-in-one solution for these statistical techniques.

Hypothesis

We hypothesize that an interactive tool can be developed using Python, leveraging libraries like NumPy for mathematical operations and Tkinter and MatPlotLib for the graphical user interface. By integrating various regression and clustering algorithms, we can create a versatile tool that simplifies data analysis tasks.

Literature Survey

Various tools and libraries provide functionalities for regression and clustering, such as Scikit-learn, R's caret package, and online platforms like Google Colab. However, these tools often require users to have a background in programming or statistics. Our tool aims to simplify the process by offering an interactive GUI that hides the underlying complexity of the algorithms used.

Methods, Data, Results, and Discussion

Methods

1. **User Interface:**
    - Developed using Tkinter to provide a graphical interface.
    - Features for adding points through textboxes, clicking on the graph, or importing CSV files.
    - Tooltips and help windows to guide users.
2. **Regression Techniques:**
    - **Linear Regression:** Implemented using the normal equation $\theta = (X^T X)^{-1} X^T y$.
    - **Polynomial Regression:** Extended linear regression to polynomial terms using the Vandermonde matrix.
3. **Clustering Techniques:**
    - **K-means Clustering:** Implemented manually to compute centroids and assign clusters, with a silhouette score to determine the optimal number of clusters.
    - **Agglomerative Clustering:** Performed using hierarchical clustering from the scipy library.
4. **Automatic Parameter Determination:**
    - Optimal polynomial degree determined using the Bayesian Information Criterion (BIC).
    - Optimal number of clusters for K-means determined using the silhouette score.

Data

We generated a couple of datasets using Scikit-learn's `make_blobs`, `make_moons`, and `make_circles` functions. Additionally, we included examples for polynomial regression.

Results

The tool successfully integrates the following features:

- Adding points through various methods.
- Performing linear and polynomial regressions with equations and Mean Squared Error (MSE) displayed.
- Conducting K-means and agglomerative clustering, with visual representations and centroid plotting.
- Saving results of regressions and clustering to files.
- Automatic determination of optimal polynomial degree and cluster number when not specified.

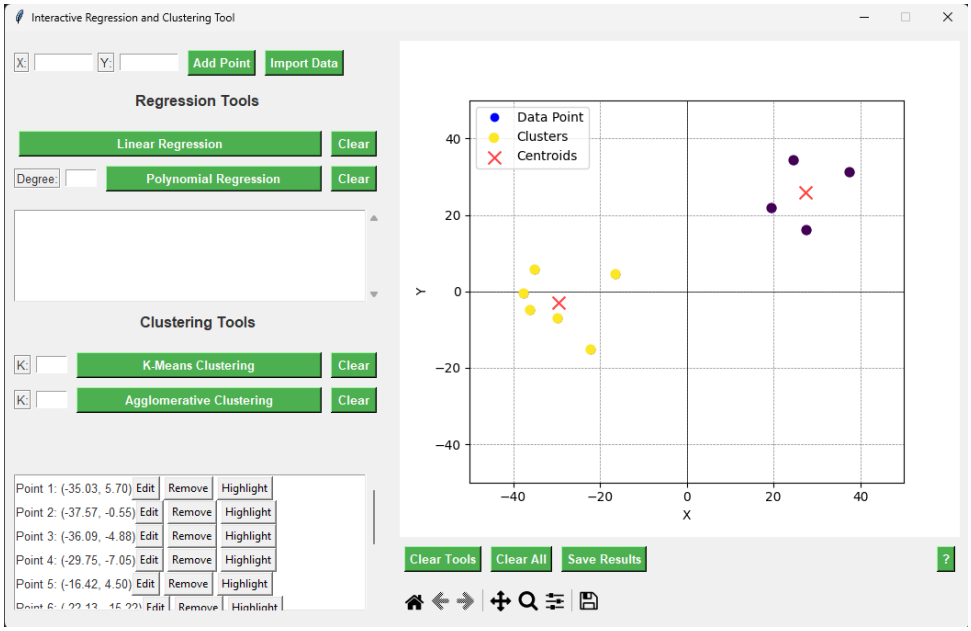

Figure 1: K-means Clustering with Centroids

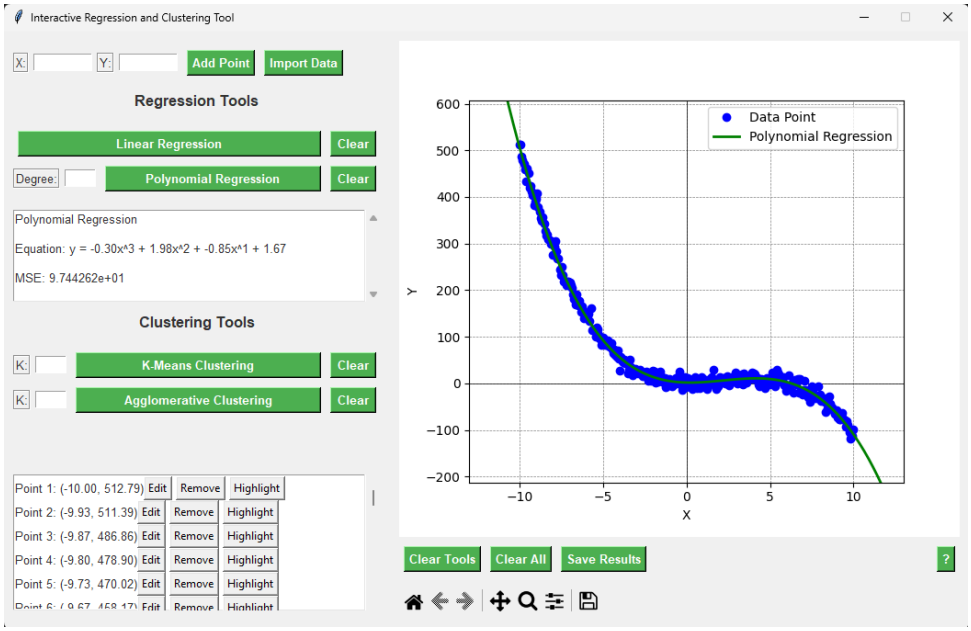

Figure 2: Polynomial Regression on a Non-linear Dataset

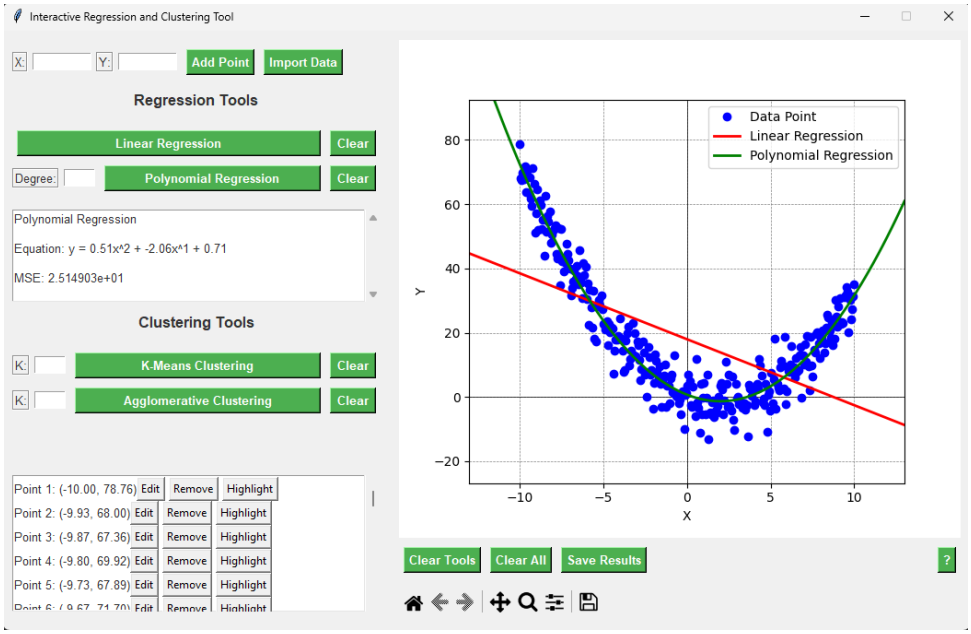

Figure 3: Linear and Polynomial Regression Comparison

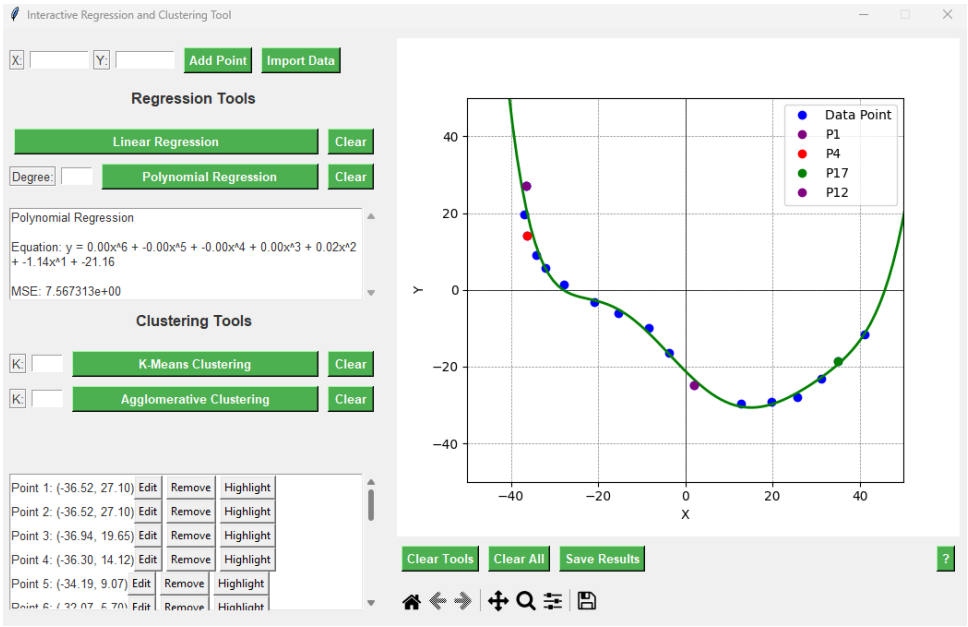

Figure 4: Highlighted Data Points

Discussion

Our tool meets the primary objectives of being user-friendly and comprehensive. The integration of various data input methods and the automatic determination of optimal parameters enhances its usability. The interactive nature of the tool allows for immediate visualization and understanding of the effects of different parameters on regression and clustering results.

This tool can be particularly useful in educational settings, such as lectures and classroom environments, where the goal is to teach regression and clustering concepts without requiring deep coding knowledge. Instructors can use the tool to demonstrate key concepts in a visual and interactive manner, making abstract statistical concepts more tangible for students. For example, the tool allows students to see how changes in polynomial degree affect the fit of a regression model or how the number of clusters influences the outcome of a clustering algorithm.

The versatility of the tool also makes it suitable for hands-on activities and assignments. Students can import their own datasets, perform various analyses, and save their results, thereby reinforcing their learning through practical application. Overall, the *Interactive Clustering and Regression Data Interface (ICARDI)* serves as a powerful educational resource, enhancing the teaching and learning experience in statistics and data analysis courses.

Features

The *Interactive Clustering and Regression Data Interface (ICARDI)* includes the following features:

- **Data Input:**
  - Add points directly by clicking on the graph.

  - Add points using textboxes for X and Y coordinates.
  - Import data from CSV files.
  - Import example datasets generated using Scikit-learn's `make_blobs`, `make_moons`, and `make_circles` functions.

- **Regression Tools:**
  - **Linear Regression:** Fit a linear model to the data and display the equation and Mean Squared Error (MSE).
  - **Polynomial Regression:** Fit a polynomial model to the data and display the equation and Mean Squared Error (MSE).
  - Automatic determination of the optimal polynomial degree using the Bayesian Information Criterion (BIC).

- **Clustering Tools:**
  - **K-means Clustering:** Cluster the data into K clusters and display the centroids.
  - **Agglomerative Clustering:** Cluster the data hierarchically and display the clusters.
  - Automatic determination of the optimal number of clusters using the silhouette score.

- **Visualization and Interaction:**
  - Display data points, regression lines, and cluster centroids on the graph.
  - Highlight specific points and add them to the legend.
  - Edit and remove points.
  - Zoom and pan the graph using the navigation toolbar.
  - Save regression results as a text file.
  - Save clustering results as a CSV file with cluster labels.

- **User Guidance:**
  - Help window explaining each UI element and its usage.

Conclusions

Interactive Clustering and Regression Data Interface (ICARDI) provides a versatile and user-friendly platform designed primarily for understanding and visualizing key concepts in regression and clustering. Its intuitive design and comprehensive features make it an invaluable educational resource, allowing users with varying levels of expertise to grasp and explore these statistical techniques effectively. ICARDI enhances the learning experience by offering immediate visual feedback and interactive exploration, making complex concepts more accessible and engaging.