CSE4/574 Spring 2024 Introduction to Machine Learning
Programming Project 1

Linear Models

Project members (Group 49):

1. Rishabh Kanodiya (rkanodiya)
2. Shreyas Waikar (swaikar)

1. Exploratory Data Analysis Report

Introduction

The initial phase of the project involved loading and exploring the California Housing dataset. This dataset is comprised of housing data for California districts, with features such as median income (MedInc), house age (HouseAge), average rooms (AveRooms), average bedrooms (AveBedrms), population, average occupancy (AveOccup), latitude, longitude, and median house value (MedHouseVal).

Data Summary

Upon loading the dataset into a DataFrame named df, the first step was to understand its structure and statistical summary. The dataset consists of 20,640 instances, each with nine features. A descriptive statistical analysis revealed the mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each feature, providing insights into the data distribution, scale, and potential outliers.

Data Quality Checks

Checks for missing values across all features indicated that the dataset is clean with no missing values, allowing for straightforward analysis without the need for data imputation.

Exploratory Data Analysis (EDA)

The EDA comprised multiple steps:

Pairwise Plots: Visualization of pairwise relationships in the dataset helped identify correlations between features and the target variable (MedHouseVal).

Correlation Matrix and Heatmap: The correlation matrix further quantified the relationships between features, with a heatmap providing a visual summary. This analysis is crucial for identifying features with strong correlations to the target variable, which can be predictors in regression models, and for detecting multicollinearity between predictors.

Histograms: The distribution of each feature and the target variable was visualized using histograms, helping to understand the skewness, and presence of outliers in the data. These insights are essential for selecting appropriate data transformations and modeling techniques.

Findings and Observations

Correlations: The correlation analysis revealed that some features have a moderate to strong correlation with the median house value, such as median income, suggesting its potential as a strong predictor in regression models.

Data Quality: The absence of missing values indicates good data quality, simplifying the pre-processing phase.

Conclusion

The exploratory data analysis provided valuable insights into the California Housing dataset's characteristics, distributions, and potential predictors for housing prices. These findings will guide the subsequent modeling phase, informing the selection of features, identification of outliers, and potential transformations needed to address skewness or scale differences. Median income emerged as a key feature for predicting housing prices based on the initial observations.

2. Linear Regression Analysis Report

Problem 2 in the project involved implementing linear regression to predict the median house value (MedHouseVal) based on median income (MedInc) in the California housing dataset. The key steps and findings from this problem are summarized below:

Dataset Preparation

The dataset comprised multiple features, but for linear regression, only the MedInc feature was used as the independent variable (X) and MedHouseVal as the dependent variable (y).

The dataset was split into training and test sets, with **80% of the data used for training** and the **remaining 20% for testing**. This division was essential for evaluating the model's performance on unseen data.

Linear Regression Model Implementation

The slope and intercept of the regression line were computed based on the training data. The calculated slope was approximately 0.419 and the intercept was approximately 0.445.

Model Evaluation

The model's performance was evaluated using several metrics:

Test Mean Absolute Error (MAE): The average of the absolute differences between the predicted and actual values, resulting in approximately 0.630.

Test Mean Squared Error (MSE): The average of the squared differences between the predicted and actual values, resulting in approximately 0.709.

Test Root Mean Squared Error (RMSE): The square root of MSE, providing a measure of the average magnitude of the error, resulting in approximately 0.842.

R^2 Score: Represents the proportion of the variance in the dependent variable that is predictable from the independent variable, resulting in approximately 0.459. An R^2 score of 1 indicates perfect prediction.

The model was visualized by plotting the regression line on the training data, which showed how well the model fits the data.

Residual Analysis

Residuals, the differences between the actual and predicted values, were plotted against the predicted values for both training and test data to assess the model's prediction errors. If the model is perfect then all points would lie on the red line as indicated in the code.

Observations and Conclusions

The linear regression model demonstrated a moderate level of accuracy in predicting the median house value based on median income, as indicated by the evaluation metrics.

The positive slope of the regression line indicated a direct relationship between median income and median house value, which aligns with intuitive expectations.

However, the R^2 score suggested that while median income is an important factor, it does not capture all the variance in median house value, pointing towards the potential benefit of including more features in the model (e.g., through multiple linear regression).

The residual plots indicated that the model might benefit from further refinement, such as incorporating additional features or exploring non-linear relationships.

In summary, the linear regression model provided a baseline for understanding the relationship between median income and median house value in the California housing dataset. While it demonstrated some predictive ability, the findings also highlighted areas for improvement and the potential benefits of more complex modeling approaches.

3. Multiple Linear Regression Analysis Report

Objective

The focus of this analysis was to explore the performance of multiple linear regression on the California housing dataset, comparing it to simple linear regression and evaluating its effectiveness in predicting median house values based on various housing features.

Dataset Overview

The California housing dataset comprises 20,640 instances, each with nine features: MedInc (median income), HouseAge, AveRooms (average rooms), AveBedrms (average bedrooms), Population, AveOccup (average occupancy), Latitude, Longitude, and MedHouseVal (median house value). The dataset contains no missing values, ensuring a smooth analysis process.

Methodology

Multiple linear regression was implemented by first enhancing the feature matrix with a column of ones to account for the intercept term. The dataset was partitioned into training (80%) and testing (20%) subsets to facilitate model training and evaluation. A closed-form solution was utilized to compute the regression coefficients, allowing predictions on the test data.

Performance Metrics

The model's effectiveness was quantified using several metrics:

Test Mean Absolute Error (MAE): 0.5332, indicating the average deviation of the predictions from the actual values.

Test Mean Squared Error (MSE): 0.5559, reflecting the average squared difference between the estimated values and the actual value.

Test Root Mean Squared Error (RMSE): 0.7456, providing a measure of the average magnitude of the errors.

R^2 Score: 0.5758, signifying that approximately 57.58% of the variance in the median house value is predictable from the features.

Observations

Comparison with Simple Linear Regression: Multiple linear regression demonstrated a noticeable improvement in prediction accuracy over simple linear regression, as evidenced by lower MAE, MSE, and RMSE values and a higher R^2 score. This suggests that including more explanatory variables helps capture the complexity of housing prices more effectively.

Feature Importance: The use of multiple features provides a more interesting understanding of how various factors collectively influence house prices, unlike simple linear regression which relies on a single predictor.

Computation Efficiency: Despite its improved accuracy, multiple linear regression maintains computational efficiency, making it a practical choice for large datasets.

Conclusion

The analysis clearly shows that multiple linear regression is superior to simple linear regression for this dataset, providing more accurate predictions with a reasonable computational cost. It effectively leverages the multidimensional nature of the housing data, making it a recommended approach for predicting median house values. The balanced performance in terms of accuracy and computational efficiency underscores its utility in real-world applications where multiple factors influence outcomes.

4. Locally Weighted Linear Regression Analysis Report

Locally Weighted Linear Regression (LWLR) is an advanced regression technique that focuses on fitting linear models to localized subsets of data, providing the flexibility to capture non-linear relationships between variables. This method is particularly useful for datasets with complex and non-linear patterns, as it adjusts the model to focus more closely on data points near the target point for prediction.

Implementation and Methodology

The implementation of LWLR in the project involved the following key steps:

1. **Data Preparation**: The dataset was split into training and test subsets using 'train_test_split' from the 'sklearn.model_selection' module. A custom function to compute Gaussian kernel weights for all data points relative to a given prediction point was developed, leveraging the numpy library for mathematical operations.

2. **Weight Calculation**: For each prediction point, Gaussian kernel weights were calculated for all training instances, based on the Euclidean distance between the training points and the prediction point. The bandwidth parameter, 'tau', played a crucial role in determining the extent of the weighting, influencing how much each point in the dataset contributed to the regression model at the prediction point.

3. **Model Fitting and Prediction**: Using the calculated weights, the LWLR model was fitted to the training data for each prediction point. This involved solving for the coefficients of the linear model in a way that accounted for the weights, emphasizing the influence of points closer to the prediction point.

4. **Parameter Tuning**: The optimal value for the `tau` parameter was determined through experimentation. Several values of `tau` were tested, and the one resulting in the best model performance, as measured by the R-squared metric, was selected. The model's performance was evaluated using a subset of the test data to make the tuning process computationally feasible.

We have found the optimal value of tau to be close to 7, but it depends on the various factors that what metric are we using (like we userd R2 score), if someone uses MSE or RMSE they may get different results. Also, how many samples are being taken into consideration is also an important factor.

Results

1. The implementation demonstrated that LWLR could adjust predictions according to local data characteristics, providing a more flexible and potentially accurate model compared to global linear regression techniques.

2. After tuning, an optimal `tau` value of 7 was identified, achieving an R-squared score of approximately 0.473 on a subset of the test data. This indicated a moderate level of explanatory power, with the model accounting for around 47.3% of the variance in the target variable.

3. The performance metrics for the model with the optimal `tau` value were as follows: Mean Absolute Error (MAE) of approximately 0.553, Mean Squared Error (MSE) of approximately 0.6356, and Root Mean Squared Error (RMSE) of approximately 0.797.

Conclusion and Recommendations

LWLR provided an in-depth approach to regression analysis, allowing for variable model complexity across different regions of the dataset. The selection of the 'tau' parameter was critical for balancing the model's bias and variance, influencing the effectiveness of the local weighting.

Despite its advantages in flexibility and local fitting, LWLR comes with **increased computational costs** compared to simple or multiple linear regression, particularly for large datasets and in the parameter tuning phase.

In summary, Locally Weighted Linear Regression offered a powerful alternative to traditional regression models by adapting to local data characteristics. Its implementation in this project showcased the potential for enhanced prediction accuracy in the presence of non-linear relationships, at the cost of increased computational complexity and the need for careful parameter tuning.

5. Model Comparison and Selection Report

Based on the results from the previous four problems in your regression analysis for predicting housing prices using the California housing dataset, here are some final recommendations and comparisons of the various approaches in terms of training and testing error:

1. **Simple Linear Regression**

- Approach: Used a single feature (MedInc) for predicting housing prices.

- Error Metrics:
- Test Mean Absolute Error (MAE): 0.6299
- Test Mean Squared Error (MSE): 0.7091
- Test Root Mean Squared Error (RMSE): 0.8421
- R^2 score: 0.4589

2. **Multiple Linear Regression**

-Approach: Utilized multiple features for predicting housing prices.
-Error Metrics:
-MAE: 0.5332
-MSE: 0.5559
-RMSE: 0.7456
-R^2 score: 0.5758

3. **Locally Weighted Linear Regression (LWLR)**

-Approach: Used a non-parametric method that weighs instances differently based on their distance from the query instance.
- Error Metrics:
- MAE: 0.576266
- MSE: 0.682739
- RMSE: 0.82628

Recommendations
Choosing the Best Setting:
Accuracy vs. Complexity: Multiple Linear Regression strikes a balance between simplicity and the ability to capture complex relationships through the inclusion of multiple input features. It provides improved prediction accuracy compared to Simple Linear Regression without the computational complexity of LWLR.

Metric for Selection: The Root Mean Squared Error (RMSE) and R^2 score are crucial for evaluating model performance. RMSE provides a measure of the average prediction error magnitude, while the R^2 score indicates how well the regression predictions approximate the real data points. In this context, a model with a lower RMSE and a higher R^2 score is preferable.

Final Recommendation: Given the results, Multiple Linear Regression is recommended for predicting housing prices with the California housing dataset. It offers a good compromise between model simplicity and predictive accuracy, showing better performance on the key metrics of RMSE and R^2 score compared to Simple Linear Regression and is less computationally intensive than LWLR.

Also, finding optimal tau in LWLR is computationally expensive and may vary from approach to approach. Like, one may use lowest RMSE or highest R2 score also the number of observations taken into account may vary hence Multiple Linear Regression looks better to us.

Application Considerations:

Dataset Characteristics: The effectiveness of a regression model can vary significantly with the characteristics of the dataset (e.g., feature correlations, non-linear relationships).

Computational Resources: For larger datasets or in scenarios where computational efficiency is a concern, the choice of model may lean towards simpler models unless the accuracy gains of more complex models justify the additional computational cost.

In summary, while the choice of regression model should be guided by the specific needs and constraints of the application, **Multiple Linear Regression** is recommended as a general approach for this dataset, offering a good balance between accuracy and computational efficiency.