Report on Linear Regression and Classification Using the Air Quality Dataset

Nazmul Hossain, 23015862

December 25, 2024

Introduction

In this report, we explore Linear Regression (subtopic 1) as a primary regression technique and Random Forest as a classification (subtopic 2) method. Using the Air Quality dataset, we applied preprocessing steps to prepare the data and implemented these models to evaluate performance. This report focuses on linear methods while briefly suggesting potential alternatives like non-linear regression models for future exploration.

Understand Subtopic 1

Linear Regression: A statistical method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data (Frost, 2019). Linear Regression is widely used for its simplicity and interpretability, making it a foundational approach in data analysis and predictive modeling (Seber & Lee, 2012).

Dataset Description

The Air Quality dataset, sourced from the UCI Machine Learning Repository (UCI Machine Learning Repository, n.d.), contains hourly data on air pollutant concentrations and meteorological variables. Key features include:

- **Pollutant Sensors**: PT08.S1(CO), PT08.S2(NMHC), PT08.S3(NOx), PT08.S4(NO2), PT08.S5(O3).
- **Meteorological Variables**: Temperature (T), Relative Humidity (RH), and Absolute Humidity (AH).
- **Target Variables**: CO(GT) for regression and Median CO Levels for classification.

Preprocessing

Key preprocessing steps included:

1. **Handling Missing Values**: Replaced missing values with column means to maintain data integrity.
2. **Numeric Conversion**: Converted numeric columns with commas to floats for compatibility and consistency.
3. **Feature Scaling**: Standardized features to zero mean and unit variance using `StandardScaler`.
4. **Polynomial Feature Expansion**: Added degree-3 polynomial features to improve Linear Regression's ability to model non-linear relationships.

Regression Analysis: Linear Regression

Linear Regression modeled CO concentrations (CO(GT)) as the target variable, assuming linear relationships between predictors and the target.

- **Results**:
  - Mean Squared Error (MSE): 2317.07
  - R-squared (R^2): 0.608
- **Discussion**: While Linear Regression achieved moderate performance, residual analysis indicated limitations in addressing complex patterns, even with polynomial features. This highlights the potential for non-linear models in future work.

Classification Analysis: Random Forest Classifier

Random Forest classified air quality levels based on derived Median CO Levels.

- **Results**:
  - Accuracy: 92.2%
  - Precision/Recall (Class 0): 94%/91%
  - Precision/Recall (Class 1): 91%/93%
- **Discussion**: Random Forest excelled in managing non-linear interactions and class imbalances, significantly outperforming baseline methods like Logistic Regression. Misclassifications in borderline cases underscored the need for further feature engineering.

Comparison and Contrast

Linear Regression and Random Forest demonstrated distinct strengths:

- **Subtopic 1: Linear Regression**: Used for predicting CO concentrations, Linear Regression is simple and interpretable but struggled with non-linear complexities despite polynomial enhancements. Future improvements could involve using non-linear regression models.
- **Subtopic 2: Classification**: Random Forest effectively classified air quality levels by managing non-linear interactions and class imbalances. Despite high accuracy, borderline misclassifications suggest the need for refined feature engineering.

Both methods relied on consistent preprocessing steps, including scaling and imputation, to ensure data quality. Future research could explore how non-linear regression models, such as XGBoost, might improve regression tasks.

Conclusion

This report highlights the utility of Linear Regression as a baseline for regression and the robustness of Random Forest in classification. The analysis underscores the importance of preprocessing and task-specific modeling approaches. Exploring non-linear regression models, such as XGBoost, could address the limitations observed in Linear Regression and further improve predictive accuracy.

References

- Frost, J. (2019). *Regression Analysis: An Intuitive Guide*. Statistics By Jim Publishing.
- Seber, G. A. F., & Lee, A. J. (2012). *Linear Regression Analysis*. John Wiley Sons.
- UCI Machine Learning Repository (n.d.). *Air Quality Dataset*. Available at: `https://archive.ics.uci.edu/ml/datasets/air+quality`