In the name of god

Assignment 1

Machine Learning

Mehrdad Baradaran

99222020

Master : Farahani

Abstract: This code is an analysis of a dataset that contains information about online news articles, including the article's content, title, and other characteristics. The goal is to predict the number of shares that an article will receive based on its features, such as the length of the title, the content, and other characteristics. This is accomplished using machine learning algorithms such as linear regression, ridge regression, and lasso regression. Additionally, hypothesis testing and correlation analysis were performed to gain insights into how the features affect the number of shares. Finally, various feature scaling techniques were tested to determine their impact on the accuracy of the regression models.

Introduction: The code reads in a dataset called "OnlineNewsPopularity.csv" that contains information about news articles. The dataset has 61 columns, including the number of words in the title and content, the number of images and videos, and the number of shares that the article received. The goal of the analysis is to predict the number of shares that an article will receive based on its features.

Methods:

1. Data Preprocessing: The first step in the analysis is to check the dataset for missing values, describe its statistical features, and visualize some of its features using pair plots and box plots. Additionally, the code creates new features by squaring the "n_tokens_title," "n_tokens_content," and "n_unique_tokens" features.
2. Correlation Analysis: The code calculates the correlation between the "shares" feature and the other features in the dataset to gain insights into which features have the most significant impact on the number of shares.
3. Hypothesis Testing: The code performs hypothesis testing to determine if there is a statistically significant difference in the number of shares between articles with short titles and long titles, positive and negative sentiment, and articles with images and those without images.
4. Regression Analysis: The code performs linear regression, ridge regression, and lasso regression to predict the number of shares

based on the features of the articles. The models' mean squared error is calculated to evaluate their accuracy.

5. Feature Scaling: The code tests various feature scaling techniques, including standard scaling, min-max scaling, robust scaling, and max absolute scaling, to determine their impact on the accuracy of the regression models.

Conclusion: The analysis of the "OnlineNewsPopularity.csv" dataset provides insights into the features that have the most significant impact on the number of shares for online news articles. The regression analysis shows that linear regression, ridge regression, and lasso regression can accurately predict the number of shares. Additionally, the analysis of the different scaling techniques shows that scaling the features can improve the accuracy of the regression models. The hypothesis testing performed in the analysis suggests that there is a statistically significant difference in the number of shares between articles with short and long titles, positive and negative sentiment, and articles with and without images.

References:

- numpy documentation: https://numpy.org/doc/
- pandas documentation: https://pandas.pydata.org/docs/
- scikit-learn documentation: https://scikit-learn.org/stable/documentation.html
- seaborn documentation: https://seaborn.pydata.org/documentation.html
- matplotlib documentation: https://matplotlib.org/stable/contents.html

The code uses several Python libraries, including NumPy, Pandas, Seaborn, SciPy, and Scikit-learn. Additionally, the dataset used in the analysis is available at https://archive.ics.uci.edu/ml/datasets/online+news+popularity.