REGRESSION METHODS FOR CUTTING-EDGE PROBLEMS

ABSTRACT

Regression methods play an important role in solving cutting-edge problems, including those in neuroscience, computer vision, and natural language processing. Although cutting-edge problems often involve large, complex datasets, regression methods can process this data into useful information. From modeling relationships between different features to predicting outcomes using supervised learning, real-life challenges can be tackled effectively. There is a wide variety of regression methods, such as linear, non-linear, and tree-based approaches. It is indispensable to recognize that each regression method has distinct features for addressing different kinds of challenges. For instance, linear regression methods can uncover hidden patterns in data and predict outcomes for unknown data. In this paper, decision tree regression, a tree-based regression method, will be employed to analyze a complex dataset consisting of brain, image, and text features. Regarding scalability, interpretability, and feasibility, regression provides a sensible approach to analyzing cutting-edge problems and uncovering hidden insights within complex datasets.

Introduction

In modern society, everything is digitalized. Challenges have been steadily increasing due to the surge of complex, high-dimensional datasets. It is undeniable that simplifying datasets has become a crucial task for addressing cutting-edge problems. Analyzing and training features can provide deeper insight into datasets and enable effective modeling of the relationships between different features.

Regression methods have always been a key tool for analyzing complex datasets, as they reveal significant relationships. Various regression techniques are available for machine learning and analysis, including linear regression, logistic regression, ridge regression, lasso regression, etc. However, decision tree regression is selected for this analysis due to its outstanding performance in handling non-linear data with complex distributions.

Related Work

In this paper, we explore a multimodal AI dataset as an example of machine learning. The EEG BraVL dataset, comprising brain, image, text, and label features, is employed for data exploration, implementation, model training, and further analysis. As emphasized in the abstract, a regression method, specifically decision tree regression, has emerged as a powerful tool to explore the EEG dataset. This tree-based regression method processes non-linear data by hierarchically splitting it into leaves based on distinct features. Ultimately, label predictions are achieved by training the data within the tree, demonstrating the effectiveness in simplifying complex datasets.

Supervised learning will be applied to the provided dataset. It is crucial to note that the configuration of decision tree regression can yield diverse results, including the number of splits, the size of leaf nodes, the number of depth searches, etc. This hierarchical tree structure aims to create a model that predicts the target variable based on the distribution of child nodes. As data is allocated into various nodes, a tree structure is constructed. Subsequently, the implementation's performance will be measured and refined, new methodology will also be discussed based on the test results. Notably, the accuracy and performance of the model will be significantly improved.

1 Project

This section will cover four sub-sections: Comprehensive Dataset Exploration, Custom Model Implementation, Result Analysis and Visualisation, Paradigm Design, and Data Splitting.

1.1 Comprehensive Dataset Exploration

1.1.1 KNOW

In this section, we aim to gain a deeper understanding of the dataset through basic statistical analysis. The EEG BraVL dataset is utilized for the implementation of machine learning. This dataset has 19417960 data items in total, with 1654 classes in the seen label category and 200 classes in the unseen label category. The brain feature data consists of 561 features, each with 16540 data. Similarly, the image feature data contains 100 features with 16540 data each and the text feature data contains 512 features with 16540 data each. The label feature data has one column with 16540 data points. The brain, image, and text data will be used in conjunction with the label data for supervised learning purposes. However, the class of the dataset will be reduced to 20 for enhancing high efficiency with easier implementation.

Given the dataset's size, exploration is critical before processing data. To uncover hidden patterns, all data are iterated to calculate statistical measures such as count, mean, std, min, 25%, 50%, 75%, max. Following this, we identify the most frequently occurring range of values based on the data range of each data frame. It is found that the most frequent brain data range is 0.0461 to 0.3733, for the image data is -3.5633 to 14.7623, and for text data is -0.6018 to -0.1584. Additionally, the peak amplitude of the brain data, representing the strength and intensity of brain signals, is identified as 9.1154, highlighting its significance in the analysis.

1.1.2 SEE

To visualize the dataset, we will use box plots, violin plots, and heatmaps for each of the features.

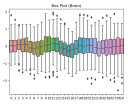

Figure 1: brain (box)

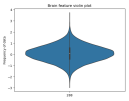

Figure 2: brain (violin)

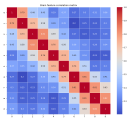

Figure 3: brain (heatmap)

Figure 4: image (box)

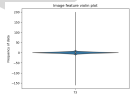

Figure 5: image

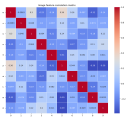

Figure 6: image(heatmap)

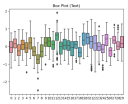

Figure 7: text (box)

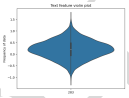

Figure 8: text (violin)

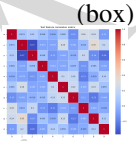

Figure 9: text (box)

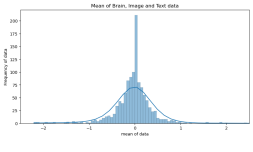

Figure 10: mean of all data (histogram)

1.1.3 FIND

The box plot illustrates the distribution of data for 30 randomly selected features from the dataset, it reveals high consistency in the features of the brain dataset, whereas the image dataset exhibits a diverse data range, and the text dataset shows a fluctuating data range. Data cleaning is necessary for both the image and the text data to achieve an optimal result.

The violin plot highlights that the distribution of the image features differs significantly from the other two feature types, exhibiting low variability with similar data points. This could suggest either a highly concentrated dataset or insufficient data for accurately representing a normal distribution. However, the latter possibility is unlikely, given that we have 100 features with 16,540 data points for each.

The heatmap emphasizes the stark difference between the brain features, with image and text features. The gradient of color in brain features indicates that data is distributed consistently, while image and text features appear less organized. This consistency makes the brain features ideal for testing the performance of the regression method.

The histogram reveals the presence of an outlier in the range of 0.1 to 0.2, where the frequency of data points significantly deviates from the expected pattern, exceeding the typical bell curve shape of a normal distribution. This observation suggests that the data does not adhere to a normal distribution, leading to inaccurate performance.

DM me for the full version