Mining Knowledge from Data Frame: A Data Analysis Report

Corresponding Author / Report Owner: Siddhant Ashwani

1st. Aayushi J Tripathy
dept. Computer Science of Engineering
Amrita School of Computing
Bengaluru, India
BL.EN.U4CSE21003@bl.students.amrita.edu

2nd S Navin Sunder,
dept. Computer Science of Engineering,
Amrita School of Computing,
Bengaluru, India
BL.EN.U4CSE21176@bl.students.amrita.edu

3rd Siddhant Ashwani
dept. Computer Science of Engineering
Amrita School of Computing
Bengaluru, India
BL.EN.U4CSE21189@bl.students.amrita.edu

***Abstract*: This report explores key concepts in machine learning and data analysis. Discussion of the importance of the observation matrix rank in the construction of classification models places an emphasis on its crucial function in linear models and addresses issues like multicollinearity and overfitting [1]. In order to better understand how to predict continuous and categorical outcomes, tasks involving regression and classification are compared to show how they differ in applications and methodology. The article also discusses methods for anticipating predictive stock prices and change percentages, showing that despite the stock market's complexity, previous data may be used to make well-informed projections.**

Keywords: data visualization, dimensionality, classification model, regression model, SVM, LDA, rank, central tendency,regularization

II. INTRODUCTION

The analysis of this report was done using the Python programming language, which includes several modules to study the report. Modules that are extensively used for assignment submission include: Numpy, which got used for all the numerical analysis; Pandas, which got utilized for all the data processing and extraction work for the data frame; and Matplotlib Matplotlib found its usage in data visualization and Statistics for finding the statistical measures of the central tendency values. The submodule utilized over here comprises Linalg, which is used to perform all the linear algebra-ated operations like the pseudo-inverse calculation and Pyplots. A thorough study of these modules was done before conducting the analysis.

II. REPORT QUESTION ANALYSIS

A. Discuss the importance of rank of an observation matrix in model building for classification.

1) Key Concepts

The greatest number of linearly independent rows or columns is referred to as Rank and is calculated using row-echelon form, gauss-seidel, and the gauss elimination process. While creating models for classification, it plays a major role in linear models involving linear algebra problems. The observation matrix, which is used to create a classification model, generally comprises feature vectors that represent the input data points, which solely depend on finding the rank of the matrix.

2) Usage Simplicity

2.1) Linear Algebra operations

Many classification problems, such as support vector machines and linear discrimanalysis, make use of the observation matrix to find the coefficient that determines the decision functions. Rank less than the number of rows or columns indicates the linear dependence between vectors, which can help us understand the multicollinearity issue and help in finding the apt relation, which can be used in the classification model.

2.2) Overfitting the curves

If the rank of the observation matrix is higher, the model is more susceptible to fitting issues. It involves such a curve, which would take up all the variations, noise, or inconsistencies that would lead to poor generalization of the curve and would reduce the accuracy of the model.

2.3) Reducing Dimensionality and Matrix inversion

High-dimensional data with a lower rank implies dependence between the vectors and helps us find out the correlation between vectors; hence, we can judge which attributes to discard involving low variance, which ultimately helps us achieve a model of greater accuracy. Along with that, less rank means the inversion of the matrix would get complicated, which would help us eradicate numerical instability and related errors.

3) Review

The rank of an observation matrix has a significant impact on a classification model's stability, accuracy, and generalizability, especially for methods or processes involving elementary linear algebra ideas. To make sure that rank-related problems are minimized and the classification model works well on both training and fresh data, it is crucial to pre-process data to handle multicollinearity, conduct dimensionality reduction where applicable, and use the right regularization approaches.

B. Discuss the regression (Ex: A2) and classification (Ex: A3) tasks. Differentiate between them.

1) Key Concepts

Regression [2], a supervised learning method, actually predicts the outcome of the data processing in terms of a continuous range of values. Finding a connection between the input feature vector and the target characteristic is the primary goal in order to provide the most accurate predictions. We just pay attention to the numerical value.

Classification, on the other hand, is a distinct kind of supervised learning in which the main goal is to divide the data values into several categories or discrete, predetermined classes in accordance with specified requirements. The output value is categorical rather than numerical, and we receive the findings in the form of labels. On the basis of a decision boundary or decision function, the model is anticipated to forecast.

2) Key Comparisons within Data

We transformed the matrix into the AX=C form in order to determine the cost of the products in the A2 question, where we had to use the pseudo-inverse method. Ultimately, we found the most precise or best-fitting X matrix value, which was simply the cost of each item in accordance with the available data on customer purchase quantity.

In contrast to the A3 question, where we had to categorize people as either affluent or poor based on how much they had paid for something, in this case, we specified clear criteria for the categorization, and the model responded to us with a tagged response.

C. Observing the stock data provided, record your suggestions to build a system that may be able to predict the price and change percentage in the future.

In order to obtain the best answer to the comparison analysis for the price vs. change percentage forecast, we employ the regression problem for the provided problem.

Although predicting stock prices is difficult due to the complexity and volatility of the Indian stock markets, we can still use old data to train our model properly. This old data spans a very long time, ranging from years, and can make our machine study the patterns and make the best possible prediction for the value of the stock price, taking into account no drastic market events.

We can examine the data to determine the likelihood that a profit occurred over a wide range of time values, study the corresponding change percentages, and use central tendency methods to determine the most frequent or average changes seen over a long period of time. With the likelihood that a loss or profit will occur, we can confidently predict future values.

Acknowledgment

I would like to express my sincere gratitude to the university, Amrita Vishwa Vidyapeetham, our lecturer, Dr. Peeta Basa Pati, and Ms. Roshni for assigning us the assignment and project and helping us with the insights through the project. Their knowledge of the subject assisted in the successful completion of this project report.

References

[1]https://aws.amazon.com/what-is/overfitting/#:~:text=Overfitting%20is%20an%20undesirable%20machine,on%20a%20known%20data%20set

[2]https://www.javatpoint.com/regression-vs-classification-in-machine-learning