 Date :

"PRACTICAL NO. 03"

AIM :

Using linear regression predict the value of continuous data.

INPUT :

Dataset containing features [independent variables] and a target variable [dependent variable] that is continuous.

OUTPUT :

A predictive model that estimates the continuous value of the target variable based on input features, as well as predictions for new data.

THEORY :

Linear regression is a statistical method used for modelling the relationship between a dependent variable and one or more independent variables. If there is a single independent variables, it is called 'simple linear regression'; if there are multiple independent variables then it is called 'multiple linear regression'. The objective of linear regression is to find the best fitting line that minimizes the error between predicted and actual values. This relationship can be expressed as :

$$y = b_0 + [b_1 \cdot x_1] + [b_2 \cdot x_2] + \cdots + [b_n \cdot x_n]$$

where,

- y is the dependent variable [target],
- x_1, x_2, x_n are independent variables,
- b_0 is the intercept,
- $b_1, b_2 \cdots b_n$ are the coefficients.

Linear regression relies on certain assumptions like, linearity, independence of errors, homoscedasticity and normally distributed errors. When these assumptions are reasonably met, linear regression provides reliable estimates.

A critical part of linear regression is the set of assumptions it makes about the data. First, it assumes a linear relationship between the dipendent and independent variables. If the relationship is non-linear, linear regression may not perform well. Second, the errors should be independent, meaning they are not correlated with eachother. Third, these errors should have constant variance so that they do not systematically increase or decrease as the values of the independent variables change. Finally, linear regression assumes that the errors follow a normal distribution.

The goodness of fitting of a linear regression model is often evaluated using the R-squared [R^2] metric, which indicates the proportion of variance in the dependent variable that is predictable from the independent variables. An R^2 value closer to 1 indicates a model that explains

a high proportions of that variance, suggesting a good fit of the data. However, high R^2 values do not guarantee a good model, as they can sometimes indicate overfitting if the model is overly complex relative to the amount of data.

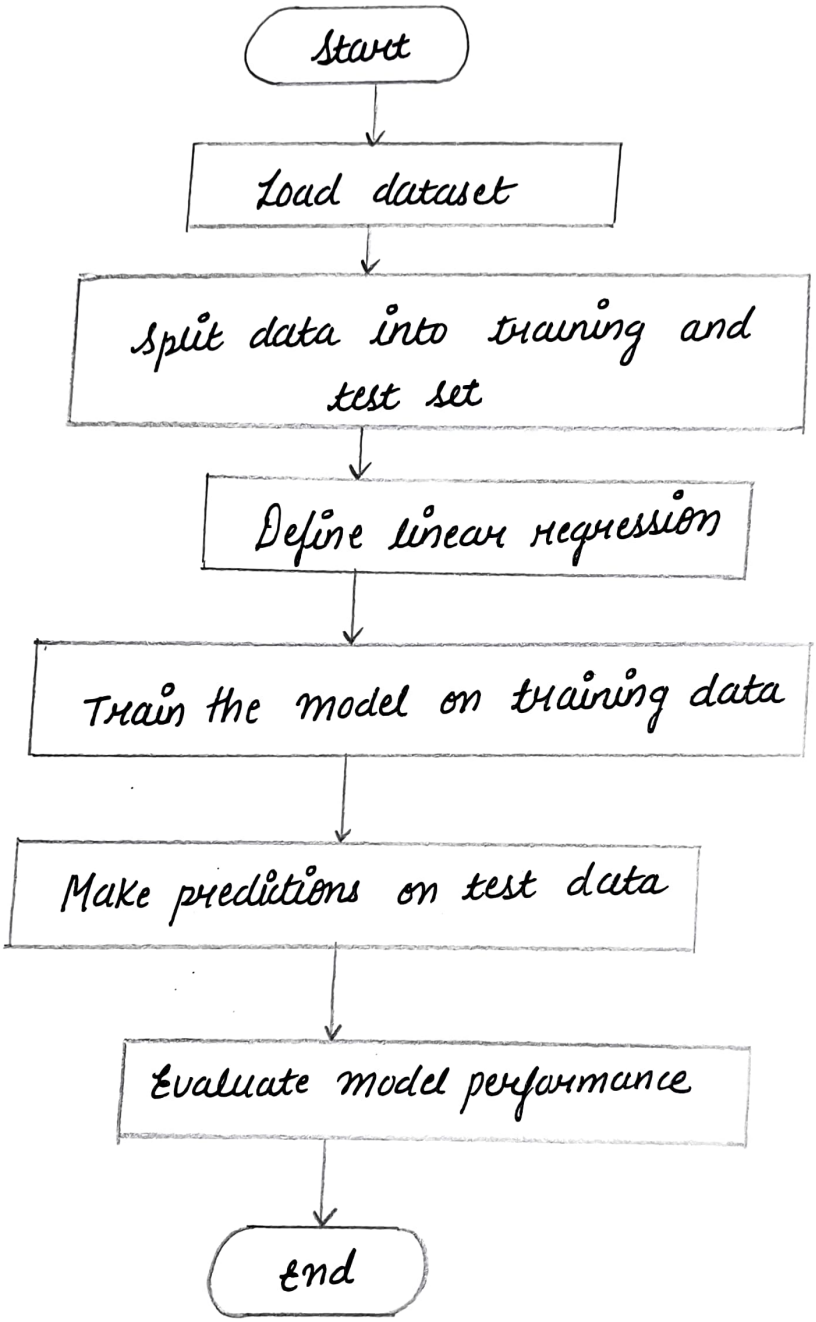
Despite its simplicity, linear regression is highly effective for many real-world applications such as predicting house prices, sales trends and economic metrics. When the assumption of linear regression are reasonably met, it provides reliable and interpretable results. However, if these assumptions are violated, alternative techniques, such as polynomial regression or regularized regression methods.

ALGORITHM :

STEP [1]: Obtain the dataset with input fectures and target variable.
STEP [2]: Handle missing data and test / training set seperation.
STEP [3]: Setup the linear regression model
STEP [4]: Use the training data to fit the linear regression.
STEP [5]: Use the trained model to predict the target variable.
STEP [6]: Evaluate the model.

CONCLUSION :

Linear regression is a simple yet powerful technique for predicting continuous data. By building a linear relationship between features and the target variable, it can make reliable predictions when the data meets certain assumptions. This method is commonly used for tasks like predicting prices, scores and other quantitative outcomes. The results depend on the quality of the input data and the degree to which assumptions of linear regression are satisfied. In practice, if assumptions are violated, alternative methods [like polynomial regression or regularized regression] may yield better results.


Start
Load dataset
Split data into training and test set
Define linear regression
Train the model on training data
Make predictions on test data
Evaluate model performance
End

"PRACTICAL NO. 04"

AIM :
Using Chi- square analysis to predict the analysis, hypothesis.

INPUT :
- The input dataset consisting of observed frequency data for categorical variable.
- Expected frequency calculated based on probabilities or derived from the dataset itself.

EXPECTED OUTPUT :
The output is a chi- square test statistic [χ^2 value] and a p- value.

THEORY :
The chi- square [χ^2] test is a non- parametric statis- -tical test used to examine the relationship between categorical variables. It assesses how observed frequencies compare to expected frequencies, given that the null hypothesis assumes no association between the variables. This method is commonly applied in tests of independence and goodness- of- fit.

1. Chi - square test of independence :

This test assests whether two categorical variables. It assesses how observed frequencies compare to expected frequencies, given that the null hypothesis assumes no association between the variables. This method is commonly applied in tests of indipendence. If the test statistics is significantly large, it implies that the variables may be associated.

2. Chi - square Goodness - of - fit test :

This test compares observed frequencies to expected frequencies under a specified distribution, typically a uniform or theoretical distribution.

The chi - square statistic formula is -

$$\chi^2 = \sum \frac{[O - E]^2}{E}$$

where, O represents observed frequency and E represents expected frequency.

This calculation is performed for each cell or category in the dataset and the values are summed to get the total chi - square statistic. A larger chi - square value suggests that the observed frequencies deviate more from the expected frequencies which could indicate that the null hypothesis is incorrect.

Degrees of freedom [df] is a crucial component in chi- square tests, as it affects the interpretation of the test statistic. For a chi- square test of independence, the degree of freedom is calculated as :

$$df = (r - 1) \times (c - 1)$$

where, 'r' is the number of rows and 'c' is the number of columns in the contingency table.

The chi- square test evaluates the null hypothesis [H_0] against an alternative hypothesis [H_1] by comparing the chi- square statistic to a critical value or by calculating the p - value. The p - value represents the probability of observing the given data assuming that H_0 is true. If the p - value is below a chosen significance level. The null hypothesis is rejected, implying that there is a statistically significant association or deviation.

ALGORITHM :

STEP [1] : Define the hypothesis whether it is null or alternative hypothesis.

STEP [2]: Collect and input the data, recording observed frequency.

STEP [3]: Calculate expected frequencies based on either a theoretical model or observed distribution proportions

STEP [4]: Apply the chi- square formula to calculate χ^2.

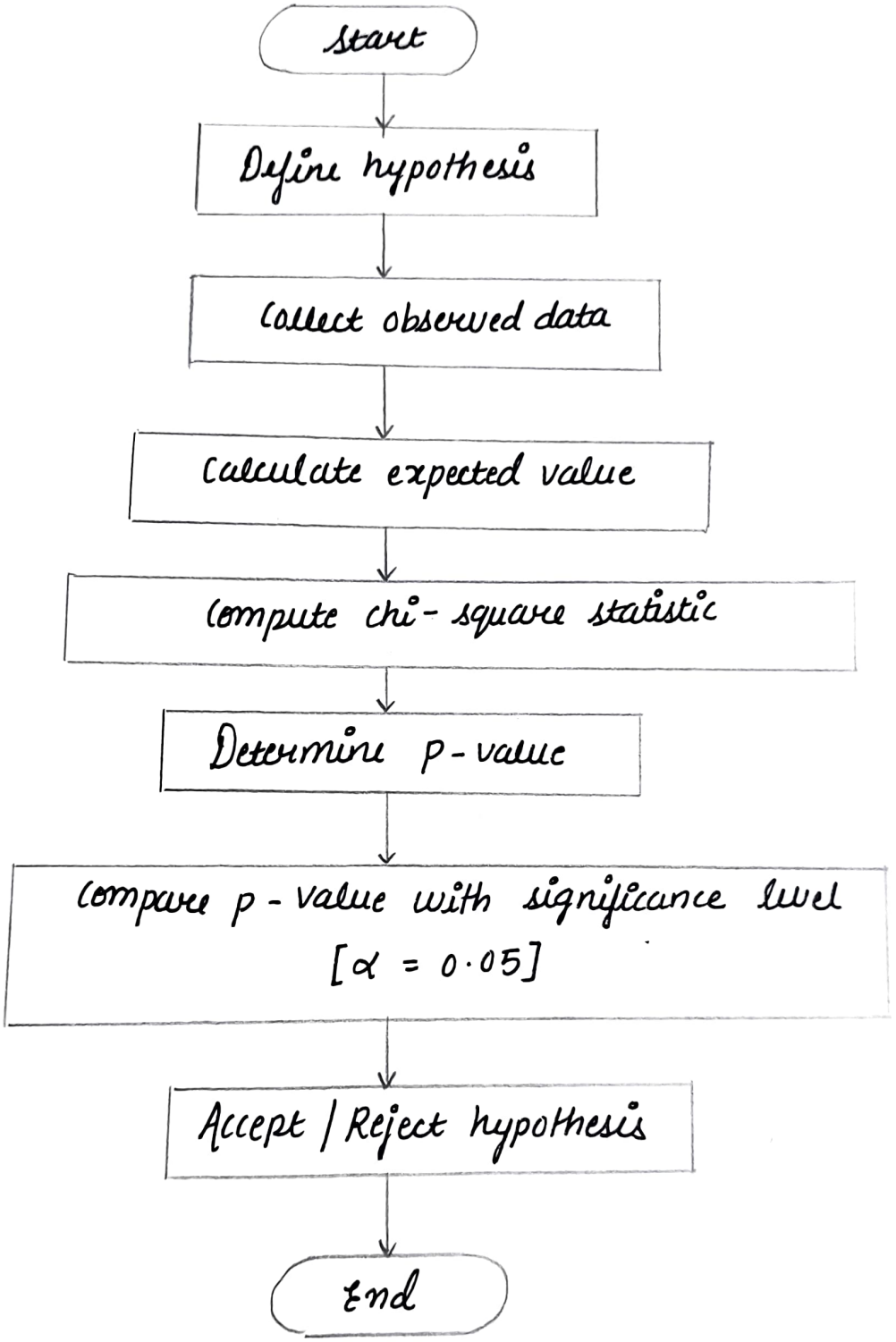
STEP [5] : Determine the p-value using the chi-square distribution with appropriate degrees of freedom.

STEP [6] : Compare the p-value with the chosen significance level to accept or reject the null hypothesis.

CONCLUSION :

The chi-square test provides a robust method for determining whether categorical variables are independent or whether an observed distribution matches an expected one. A significant chi-square statistic indicates that the observed frequencies differ from what is expected under the null hypothesis, leading to a rejection of H_0. This test is widely applicable in research fields like social science, marketing and biology where understanding relationships in categorical data is essential.

Institute of Technology- Polytechnic
Adgaon, Nashik- 422 003. Date :



 Date :

"PRACTICAL NO. 05"

AIM :

Predict the e-mail spam or ham [not spam] using classification algorithm.

INPUTS :

A dataset of labelled e-mails, with each e-mail marked as 'spam' or 'ham'. Features include the e-mail text content and metadata like subject line and sender information.

EXPECTED OUTPUT :

A trained classification model that predicts whether an e-mail is 'spam' or 'ham'. It can label new, unseen e-mails on learned patterns with high accuracy.

THEORY :

E-mail spam classification is a binary classification problem, where e-mails are labelled as either 'spam' or 'ham'. Machine learning algorithms, such as Naive Bayes, Support Vector machine [SVM] and logistic Regression are commonly used for this task. The Naive Bayes classification is effective due to its assumption of independence among features which works well in text-based application. In contrast, SVM tries to find the optimal hyperplane that seperates 'spam' and 'ham' e-mails.

Text data is transformed into a numerical form through vectorization techniques like TF-IDF which gives importance to words that are unique to certain categories. This model helps identify distinctive spam words, such as "offer" or "click" that are less frequent in ham e-mails.

One of the most popular algorithms for spam classification is the Naive Bayes classifier. It is well suited to text classification task because it operates on the principle of word independence, where the presence of one word is considered independent of the presence of another. Despite this assumption being somewhat simplicit, it performs exceptionally well in spam detection. The algorithm calculates the probability of an e-mail being spam, given the occurence of specific words.

Another commonly used algorithm in spam detection is the support vector Machine [SVM] is a supervised learning algorithm that classifies data by finding the optimal hyperplane that maximizes the margin between the classes. In the context of spam classification, SVM identifies the boundary that best seperates spam from ham e-mails. SVM is particularly effectively in high-dimensional spaces, which is beneficial in text classification, where each unique word is a dimension.

Model performance in spam detection is evaluated using metrics such as accuracy, precision, recall and F1 score. Accuracy represents the proportion of correctly classified e-mails; while precision and recall focus on the

performance related to spam detection. Precision measures how many e-mails predicted as spam and are indeed spam, while recall measures the proportions of actual spam e-mails correctly classified. The F1 score combines precision and recall to give a balanced view of model performance, especially when handling imbalanced data, which is common in spam classification as spam is usually a smaller fraction of overall e-mails.

ALGORITHM :

STEP [1] : Obtain a labelled dataset of spam and ham e-mails
STEP [2] : Cleaning of the text data, tokenization and vectorization.
STEP [3] : Choose a classification algorithm [e.g. Naive Bayes, SVM]
STEP [4] : Train the model on pre-processed data.
STEP [5] : Evaluate the model on a test set using metrics like accuracy, precision, recall and F1 score.
STEP [6] : Use the trained model to classify new e-mails as spam or ham.

CONCLUSION :

In this project, we successfully developed a machine learning model that can distinguish between spam and ham e-mails with a high degree and accuracy. The classification algorithm utilized text-based features, enabling it to

recognize common spam indicators, enhancing e-mail filtering capabilities. Such a model is beneficial for improving user experience and protecting against unwanted or harmful e-mails.

Institute of Technology- Polytechnic

Adgaon, Nashik- 422 003. Date :