classmate
Date
Page

AI Assignment - 4

Logistic Regression

Aim: Write a program to implement a logistic regression for a given dataset eg. titanic

Alg: Logistic Regression
Platform: Arch linux x86_64
IDE : VS code

(*) **Theory**

→ What do you mean by logistic regression?

It is a statistical method for predicting probability of a binary outcome. Its commonly used when the dependent variable is categorical, meaning it has only possible outcomes like "yes" or "no", 0 or 1, or true or False.

(*) What are the different types of logistic regression?

(1) Binary logistic Regression: This is the standard form of logistic regression; used when the dependent variable has only two possible outcomes, for eg whether a patient has some disease or not.

2. Multilingual Logistic Regression :

This type is used when the dependent variable has three or more unordered categories. For instance, predicting the type of flower based on its characteristics where the outcome can be one of several possible flower types.

3. Ordinal Logistic Regression : This form is applied when the dependent variable has three or more ordered categories. It's suitable for scenarios where there is a natural ordering between the categories.

(*) The Titanic Dataset :

It is a well known and frequently used dataset in the field of machine learning and data science.

It contains information about passengers aboard the ill fated maiden voyage taking titanic in April 19

(*) FAQs

(1) How do we handle categorical variables in logistic regression?

→ Handling categorical variables in logistic regression is an important aspect of building accurate prediction models. Categorical variables represent non-numeric data such as names, labels or categories.

(A) One Hot encoding:

For a categorical variable with 'n' unique categories, we create n binary columns, one for each category.

(B) Label Encoding: Label encoding assigns a unique integer to each category. You need to be cautious with label encoding for logistic regression as it may imply a linear relationship between the categories that may not exist.

(C) Frequency (count) Encoding:

It replaces each category with the count of its occurrence in the data set.

(2) Explain different feature types in logistic regression.

→ In logistic regression, the choice of features (Independent variables) is a critical aspect of model development.

(A) Continuous (Numerical) Features:

They are numeric variables that can take any real value, within a range. Examples

include age, income, temperature and weight.

(B) Categorical Features

They rep. discreet categories or labels, examples include gender, color, country, products type etc.

(C) Binary Features

They are a subset of categorical features with exactly 2 categories. Examples include gender (male/female), yes/no responses, true false values.

Q(3) What are assumptions made in logistic regression?

→ Logistic Regression, like any other stat model is based on a set of assumptions, that should be considered and validated to ensure the model's reliability and interpretability.

(1) Independence of Observations

Each observation must be independent of each other. The probability of one's observational outcome should not affect the other.

(2) No endogeneity

Endogeneity refers to a situation where an independent variable is correlated with the error term.

This can lead to biased coefficient estimates.

(3) Sufficient Sample Size:

Logistic regression assumes a sufficiently large sample size to provide reliable parameter estimates. Small sample sizes can lead to unstable estimates and may not adequately represent the situation/population.

(4) Absence of outliers:

Outliers have a substantial effect on logistic regression results; particularly when the model is sensitive to extreme observations.

(Q.4) Can we solve multiclass classification problems using regression, if yes, then how?

→ Yes. While logistic regression is originally designed for binary classification tasks, there are several approaches to adapt it for multiclass classification.

(1) Multinomial Logistic Regression (Softmax Regression)

→ multinomial logistic regression, is a direct extention of binary logistic regression to handle multiclass problems.

(2) Ordinal logistic Regression:

→ Ordinal logistical regression can be used for multiclass problems with ordered categories.

→ It is suitable when the classes have a neutral rank of order.

(3) Regularized Logistic Regression

→ Regularized logistic regression methods like L1 (lasso) and L2 (Ridge) Regularization can be applied to multiclass problems to prevent overfitting and select important features.

Q.5 # What insights can be gained from the data analysis and models predictions?

→ (1) Identification of key factors:

Data analysis can reveal which variables or features have the most significant influence on the outcome or target variable

(2) Pattern Discovery:

Data analysis can uncover hidden patterns or trends in the data, such as seasonal effects cyclical behaviours, or co-relation between variables.

③ Segmentation :

Analysing data can lead to the discovery of distinct customer segments or subgroups with unique behaviours, preferences and needs

④ Feature Importance :

Models often provide information on importance of different features in making predictions.