29th July 2024
Monday

Statistical Analysis & Modelling
FINAL EXAM

UJWAL.P
V01107757

Section - A - [PART - A]

(a) A classification problem is a type of predictive modeling task where the output variable is a category or class label. The goal is to assign input data into one of several predefined classes.
They are evaluated using metrics like accuracy and precision.
whereas Regression problem predicts a continuous numeric value, such as house prices, and is evaluated using metrics like mean squared error and R-squared.

Key Differences

- Output type : classification predicts discrete class labels (eg: spam or not spam, fraud or not fraud) whereas regression predicts continuous numeric values (eg: price, temperature).
- Evaluation Metrics : classification is evaluated using metrics like accuracy, precision, recall, F1-score, ROC-AUC whereas regression uses metrics like mean squared error (MSE), root mean squared error (RMSE), and R-squared (R^2).

Three Algorithms for Classification
(i) Decision Trees
(ii) Support Vector Machines (SVM)
(iii) k-Nearest Neighbors (k-NN)

(b) In Logistic Regression, the 'odds ratio' represents the ratio of the odds of an event occurring to the odds of it not occurring. It quantifies how the odds change with a one-unit change in the predictor variable.

The relation to coefficients is that the odds ratio is obtained by exponentiating the logistic regression coefficient (e^{β_i}). If β_i is the coefficient of the predictor X_i, then the odds ratio is e^{β_i}. This means that for a one-unit increase in X_i, the odds of the outcome occurring are multiplied by e^{β_i}.

(c) Principal component Analysis (PCA)

PCA is a dimensionality reduction technique used to reduce the number of variables in a dataset while retaining as much variability as possible.

It transforms the original variables into a new set of uncorrelated variables called principal components, which are ordered by the amount of variance they capture from the data.

Application in Business Analytics:

PCA is used for data visualization, noise reduction, and identifying key variables in large datasets, which simplifies modeling and improves performance

Section - B - [PART - A]

(a) A Time series problem involves predicting future values based on previously observed values over time, while a regression problem predicts a continuous outcome based on input variables without necessarily considering the order or sequence of data. The test-train split process in Time series differs in that the data is split chronologically, ensuring that training data precedes test data to preserve temporal order and prevent data leakage, whereas in typically regression problems, data can be split randomly.

(b) Stationarity in time series data means that the statistical properties, such as mean, variance, and autocorrelation are constant over time. It is crucial for time series modelling because many forecasting methods assume stationarity. To check for stationarity, visual inspection of plots, statistical tests like the Augmented Dickey-Fuller (ADF) test, and analyzing autocorrelation functions are commonly used. The ADF test is a popular test to determine if a time series is stationary.

(c) In time series modeling, data objects are often formatted to a standard date-time format. If the data is in DD-MM-YYYY format, it can be converted to a datetime object in Python using pd.to_datetime with the format parameter like
pd.to_datetime(date_string, format='%d-%m-%Y').
common evaluation metrics for time series models include Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics help assess the accuracy and performance of time series forecasts.