Machine Learning Workshop 1 – CSE 2793

MINOR ASSIGNMENT-8: REGRESSION MODELS IN ML

Objectives

1. Understand the concept and application of regression analysis.
2. Implement different regression models.
3. Evaluate the performance of the models using various metrics.

Dataset

Insurance Charges Dataset dataset includes information about medical charges billed by health insurance companies, with features like age, sex, BMI, children, smoker status, region, and the charges billed.

Link - https://www.kaggle.com/datasets/simranjain17/insurance?select=insurance.csv

Tasks

The task is to predict the amount to insurance charges to be paid as a function of the other features.

Task 1: Data Exploration and Preprocessing

1. Load the dataset and display the first few rows.
2. Perform basic statistical analysis to understand the distribution of the features.
3. Check for missing values and handle them appropriately.
4. Check for categorical features and convert them to numerical features.
5. Perform feature engineering, including the creation of new features and scaling of numerical features.
6. Split the data into training and testing sets.

Task 2: Implement Regression Models

1. Train the following regression models:
   - Linear Regression
   - Decision Tree Regression
   - Random Forest Regression
   - Gradient Boosting Regression
   - Support Vector Regression (SVR)
2. For each model, train it using the training set and predict on the testing set.

Task 3: Model Evaluation

1. Evaluate each model using the following metrics:
   - Mean Absolute Error (MAE)
   - Mean Squared Error (MSE)
   - Root Mean Squared Error (RMSE)
   - Mean Absolute Percentage Error (MAPE)
   - R-squared (R2)
2. Compare the performance of the models based on these metrics and find out which model performs the best.

Deliverables

1. A Jupyter notebook containing:
   - Data exploration and preprocessing steps.
   - Implementation of the regression models.
   - Evaluation of the models.

2. A report summarizing your findings and comparing the performance of the models.

Machine Learning Workshop 1 – CSE 2793

MINOR ASSIGNMENT-9: BINARY CLASSIFICATION

Objectives

1. Understand the concept and application of binary classification.
2. Implement different binary classification algorithms.
3. Evaluate the performance of the models using various metrics.

Dataset

Download the Employee dataset from Kaggle.

Link - https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset

This dataset contains information about employees in a company, including their educational backgrounds, work history, demographics, and employment-related factors. It has been anonymized to protect privacy while still providing valuable insights into the workforce.

This dataset can be used for various HR and workforce-related analyses, including employee retention, salary structure assessments, diversity and inclusion studies, and leave pattern analyses. Researchers, data analysts, and HR professionals can gain valuable insights from this dataset.

Tasks

The task here is to find whether an employee is going to continue or leave the organization.

Task 1: Data Exploration and Preprocessing

1. Load the dataset and display the first few rows.
2. Perform basic statistical analysis to understand the distribution of the features.
3. Perform different visual exploratory data analysis such as
   i. Histograms
   ii. Correlations
   iii. Pair wise plots
   iv. Box plots
4. Check for missing values and handle them appropriately.
5. Check for outliers and handle them appropriately.
6. Check whether the dataset is balanced or not.
7. Split the data into training and testing sets.

Task 2: Implement Binary Classification Models

1. Train the following classification models:
   - Logistic Regression
   - Decision Tree
   - Random Forest
   - Support Vector Machine (SVM)
   - K-Nearest Neighbors (KNN)

2. For each model, train it using the training set and predict on the testing set.

Task 3: Model Evaluation

1. Evaluate each model using the following metrics:
   - Accuracy
   - Precision
   - Recall
   - F1 Score
   - ROC-AUC Score
2. Plot the ROC curve for all models in a single graph.
3. Compare the performances of the models based on these metrics and find out which model performs best for this task.

Deliverables

1. A Jupyter notebook containing:
   - Data exploration and preprocessing steps.
   - Implementation of the binary classification models.
   - Evaluation of the models.
2. A report summarizing your findings and comparing the performance of the models.

Machine Learning Workshop 1 – CSE 2793

MINOR ASSIGNMENT-10: PRINCIPAL COMPONENT ANALYSIS

Objectives

- Understand the concept and application of PCA.
- Implement PCA for dimensionality reduction.
- Visualize and interpret the results of PCA.

Dataset

Use the Wine dataset available from the UCI Machine Learning Repository. This dataset contains the chemical analysis of wines grown in a particular region of Italy, which are classified into three different cultivars.

Tasks

Task 1: Data Exploration and Preprocessing

- Load the dataset and display the first few rows.
- Perform basic statistical analysis to understand the distribution of the features.
- Check for missing values and handle them appropriately.
- Standardize the features if necessary.

Task 2: Implement PCA

- Perform PCA on the standardized dataset to reduce dimensionality.
- Determine the number of principal components to retain by analyzing the explained variance ratio.

Task 3: Visualization of Principal Components

- Visualize the data in the new principal component space using scatter plots.
- Color-code the scatter plots by the wine cultivars to see if the PCA helps in distinguishing between the classes.

Task 4: Interpretation of Results

- Analyze the loadings (coefficients) of the original features on the principal components.
- Discuss how the principal components can be interpreted based on the loadings.

Task 5: Classification Using Principal Components

- Use the principal components as features to train a classification model (e.g., logistic regression, decision tree).
- Evaluate the classification performance and compare it with the performance using the original features.

Deliverables

- A Jupyter notebook containing:
- Data exploration and preprocessing steps.

- Implementation of PCA.
- Visualization of the principal components.
- Interpretation of the PCA results.
- Implementation and evaluation of a classification model using principal components.