PRACTICAL RECORD

Name : Sahil. D. Marbate.

College Name : GCOEN.

Year : 3yr / 6th sem

Subject : Professional Skills -

PAGE NO.:
DATE : / / 20

PRACTICAL - 4

Aim: Correlation and Covariance

(a). find the correlation matrix.

(b). Plot the correlation plot on dataset and visualize giving an overview of relationships among data on iris data.

(c). Analysis of covariance: variance (ANOVA), if data have categorial variables on iris data.

Theory:

Correlation: Correlation measures the strength and direction of the linear relation-ship between two variables. It ranges from -1 to 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

Covariance: Covariance measures the degree to which two variables change together. If the covariance is positive, it indicates that as one variable increases, the other variable tends to increase. If it is negative, it indicates that as one variable increases, the other variable tends to decrease.

Pairplot: A pairplot is a graphical representation of relationships between pairs of variables in a dataset.

ANOVA is a statistical method used to compare the means of three or more samples to determine if they are significantly

Teacher's Signature _____

different from each other.
Some of the methods are: one-way ANOVA, F-statistics and P-value.

Here we are performing ANOVA to text whether there is a significant difference in the 'sepal length' variable among the different species of iris flower ('setosa', 'versicolor', 'virginica').

- Inputs: filename : iris.csv
  having the attributes: sepal_length, sepal_width, petal_length, petal_width, species.

- Libraries used: pandas, numpy, seaborn, matplotlib, scipy, sklearn.

- Conclusion: De program to implement the correlation and convariance, plotting correlation plot on dataset, analysis of convariance (ANOVA) on iris dataset is performed successfully.

Ashakar
8/12/2024 (A)

PAGE NO.:
DATE.: / / 20

PRACTICAL No. 5

Aim: REGRESSION MODEL

Import a data from web storage. Name the dataset and now do logistic Regression to find out relation between variables that are affecting the admission of a student in an institute based on the his or her GRE score, GPA obtained and rank of the student. Also check the model is fit or not. Require (Foreign), require (MASS).

Theory:

In data analysis and machine learning projects, acquiring data is often the first step. While local dataset are common, something the required data resides on the web.

logistic regression is a statistical technique used for modelling the probability of a binary outcome based on one or more predictor variables. In this practical, we aim to analyze the factory affecting student admission in an institute using logistic regression. we will import a dataset from web storage, perform logistic regression analysis, and evaluate the model's fit.

It is a regression model used when the dependent variable is binary. It estimates the probability that a given instance belongs to a particular category. In our case, the binary outcome variable is admission.
(1 for admitted, 0 for not admitted)
steps include importing the Dataset, Data preprocessing, Model fitting, Model evaluation, Interpretation of results.

Teacher's Signature _____

PAGE NO.:
DATE.: / / 20

- Inputs : filename :
  Attributes are: admit, gre, gpa, rank

- Labraries used:
  (a). pandas
  (b). numpy
  (c). statsmodels
  (d). scipy

- Conclusion:

Logistic regression is a powerful tool for analyzing the factors influencing binary outcomes, such as student admission. By performing logistic regression analysis, we can identify significant predictors and develop insights into a admission process. This practical demonstrates the institutions using or as can application of logistic regression successfully.

Thakare
15/2/2024 (A)

PRACTICAL No. 6

Aim: MULTIPLE REGRESSION MODEL

Apply multiple regressions, if data have a continuous Independent variable. Apply on dataset.

Theory:

Multiple linear Regression is a statistical method used to analyze the relationship between two or more independent variables and a single dependent variable. It extends the simple linear regression model, which only considers one independent variable.

- Key ~~variab~~ Concepts:
  1. Dependent Variable
  2. Independent Variables ($X_1, X_2, \ldots X_n$)
  3. Coefficients ($B_0, B_1, \ldots B_n$)
  4. Intercept (B_0)

- Multiple linear regression relies on several assuptions, including: Linearity, Independence, Homoscedarticity, Normality, No multicollinearity.

After fitting the regression model, we typically evaluate its performance using metrics such as,

1. Mean squared Error (MSE)
2. R-squared (R^2)

- Inputs:

filename: california_housing.csv.

PAGE NO.:
DATE.: / / 20

- Libraries used:
  (1) Pandas
  (2) Sklearn

- Algorithm:
  (i) Load the required python packages using import.
  (ii) Load the 'california_housing.csv' dataset using read_csv() function of panda library.
  (iii) Describe the dataset using describe() and info() function.
  (iv) Print the correlation matrix using corr() function.
  (v) Create the required model and train them.
  (vi) Uses the test case to predict using the model.
  (vii) Print the required results and figures.

Conclusion:
We have successfully created and executed a program which applies multiple regression on the given dataset.

Thakare
7/3/2024

Amar KRISH

Teacher's Signature

PRACTICAL No. 7

Aim: CLASSIFICATION MODEL

(a) Install relevant package for classification.
(b) Choose classifier for classification problem.
(c) Evaluate the performance of classifier.

Theory:

Classification is a supervised learning task where the goal is to predict the where the goal is to predict the categorial class labels of new instances based on past observations. In other words, it involves assigning instances to one of several predefined classes or categories.

Key concepts:

(1) Classifier
(2) Features (X)
(3) Target variable (Y)
(4) Training and Testing Data.

Once the classifier is trained and tested, its performances is evaluated using various matrics including:

(1) Accuracy.
(2) Confusion matrix
(3) Precision, Recall and F-1 score

Support Vector machine (SVM) is a powerful supervised learning algorithm used for classification tasks.
It finds the optimal hyper plane that best separates the classes in the features space.

- Inputs:

  Filename : iris.csv

- Libraries used:

  (1). Pandas

  (2). sklearn.datasets

  (3). sklearn.svm

  (4). sklearn.metrics

  (5). sklearn.linear_model

  (6) sklearn.model_selection

Conclusion:

We have successfully created and executed a program which installs relevent packages for classification, choose classifier for classification problem and evaluates the performance of classifier.

Thakur
14/3/2024

PRACTICAL No. 8

Date: / /

Aim: CLUSTERING MODEL

(a) Clustering algorithms for unsupervised classification.
(b) Plot the cluster using matplotlib visualizations.

Theory:

Clustering Algorithms for unsupervised classification: Clustering algorithms are a fundamental part of unsupervised machine learning used for grouping similar data points together. One of the most commonly used clustering algorithms is k-means. k means is a partition-based clustering algorithm that divides the dataset into 'k' distinct, non-overlapping clusters.

Plotting Cluster Data using Matplotlib:
Matplotlib is a popular python library used for creating static, interactive and animated visualizations in Python. In the context of clustering matplotlib is often used to visualize the clustered data points and the cluster centers.

Practical Implementation:
(1. Data Generation
(2. k means Clustering
(3. Fitting the Modal
(4. Visualizations
(5. Display

Libraries Used:
(1. matplotlib
(2. sklearn.cluster
(3. sklearn.datasets

Teacher's Signature

Page No:
Date: / /

Inputs: dataset make_blobs
- This dataset is imported from the 'Sklearn.datasets' module.
- Isotropic gaussian blobs for clustering.

Algorithm:
(1) Choose the number of clusters k Randomly initialize k cluster centroids $c_1, c_2, \ldots c_n$.
(2) For each data point x_i, calculate distance to each centroid c_j using distance matrixs
(3) Assign x_i to cluster with nearest centroid arg min$_j$ {distance (x_i, c_j)}
(4) For each cluster calculate new centroid c_j as the mean of all data points assigned to cluster j.
(5) Check if the centroids have converged. (i) The centroids on longer change significantly between iterations. (ii) Maximum number of iteration is reached.
(6) Once convergence is achieved, each data point is assigned to a cluster, and the centroids represent the final cluster centers.

Conclusion: The combination of k-means clustering and matplotlib visualization allows us to effectively group data points into cluster and visually inspect the clustering results. This technique is valuable for various applications including outcomes segmentations, anomaly detection and pattern recognition among others.

21/3/2024

PRACTICAL No. 9

Aim: Write a program to implement k-Nearest Neighbour algorithm to classify the iris dataset Print both correct and incorrect predictions.

Theory:

(1). **Iris Dataset:**

The Iris dataset is a classic dataset in machine learning and statistics, containing measurements for 150 iris flowers from three different species: setosa, Versicolar and Virginica. Each sample consists the four features: sepal length, sepal width, petal length and petal width.

(2). **k-Nearest neighbors (k-NN) algorithm:**

The k-nearest neighbor algorithm is a simple yet effective supervised learning algorithm used for classification and regression tasks.

For classification, the algorithm predicts the class of a new data point by considering the majority class among its k nearest neighbors, where k is a hyper parameter set by the user.

(3). **Practical Implementation.**

(a). Loading the Dataset
(b). Splitting the Data
(c). Creating the k-NN classifier.
(d). Training the model.
(e). Making predictions.
(f). Evaluating predictions.

Inputs: filename : iris.csv.

Algorithm:

(1). load the dataset containing features (x) and target label (y)
(2). Split the data into training and testing sets. Typically, 80 of the data.
(3). Choose the number of neighbors (k) for the KNN algorithm.
(4). Intialize the KNN classifier with closer k value, than the model using the training data (x_train, y_train) using the 'fit' method.
(5). Calculate the distance between data point and all training data points.
(6). selects the k nearest neighbors based on calculated distance.
(7). Determine the class label of data point based on majority class k nearest neighbors.
(8). Store predicted class label for the data point.
(9). Compose the predicted labels with the actual labels from the set (y_test).
(10). Calculate the accuracy of the model by dividing the number of correct predictions by total number of predictions.
(11). Print the accuracy, both correct and incorrect predictions, along with the corresponding actual and predicted labels.

Conclusion:

The practical demonstrate the implementation of the KNN algorithm for classification task using the Iris dataset. Successfully. The K-NN algorithm performs well on the Iris dataset achieving high accuracy in predicting the species of the iris flower based on the their feature measurements.

(17)

Aparna
4/4/2024

PRACTICAL No. 10

Aim: Implement k-mean algorithm to classify the iris data set.

Theory:

K-means Algorithm:

K-Means is an iterative clustering algorithm that partitions data into k clusters. It aims to minimize the within-cluster sum of squares (WCSS) by iteratively assigning data points to the nearest cluster centroid and updating the centroids until convergence

(1) **Initialization:**
k initial centroids are randomly chosen from the data points or generated using a specific initialization technique like k-means++.

(2) **Assignment step:** Each data point is assigned to the nearest centroid based on a distance metric, commonly the Euclidean distance.

(3) **Update step:** After all data points have been assigned, the centroids are recalculated as the mean of all data points assigned to each cluster.

(4) **Convergence:** steps 2 and 3 are repeated until the centroids no longer change significantly or a maximum number of iterations is reached.

Page No:
Date: / /

- Inputs: Iris.csv

- Libraries used:
  (1). sklearn (scikit-learn)
  (2). matplotlib

Algorithm:

(1). Importing libraries: The neccessary libraries are imported, including pandas, train_split for data splitting, etc.
(2). The iris dataset is loaded using 'pd.read_csv' from the specified file path.
(3). The features (x) and target (y) are separated from the dataset. In this case, x contains all columns except the last one (features), and y contains the last column.
(4). The dataset is split into training and testing sets using 'train_test_split'.
(5). A k-means clustering model is initialized with 3 clusters (k=3) and fitted the training data using 'fit'.
(6). The model predicts cluster labels for the test data using 'predict'.
(7). The accuracy of the clustering model is calculated using 'accuracy_score' by comparing the predicted labels with the actual labels from the test sets.
(8). The accuracy score, dataframe created to display the actual and predicted labels for each sample in the test set are displayed.

Teacher's Signature

Conclusion:

K-means algorithm to classify the iris dataset is implemented successfully. The k-means algorithm provides a powerful tool for clustering and exploratory data analysis. We can uncover patterns and relationships among iris species based on their morphological characteristics.

A

Shaikat
10/4/2024

GOVERNMENT COLLEGE OF ENGINEERING

Sector- 27, Mihan Rehabilitation Colony Khapri, Nagpur-441 108 (Maharashtra)

☎ Phone No. 07103&295226(P) 295220(O) Website : www.gcoen.ac.in

E-mail : principal.gcoenagpur@dtemaharashtra.gov.in

Subject: Professional Skills-II Lab

List of Practical's

Sr. No.	Aim of the Practical
1	Perform **DESCRIPTIVE STATISTICS** using summary, str, quartile function on mtcars & cars datasets.
2	Implement **READING AND WRITING DIFFERENT TYPES OF DATASETS** (.txt, .csv) from Web and specific disk location
3	Implement **VISUALIZATIONS of** data distributions using box, scatter plot and outliers using histogram, bar chart and pie chart plot.
4	Find **CORRELATION** matrix, plot **COVARIANCE** and Analysis of covariance: variance (ANOVA) on iris dataset.
5	Implement **REGRESSION MODEL** using Logistic Regression.
6	Implement **MULTIPLE REGRESSION MODEL** using multiple regressions.
7	Implement **CLASSIFICATION MODEL** and evaluate the performance of classifier.
8	Implement **CLUSTERING MODEL** and plot the cluster.
9	Implement **k-Nearest Neighbour** algorithm to classify the iris data set.
10	Implement **k-means** algorithm to classify the iris data set.

Roma Goel

Roma Goel &

Thakare

Monali Thakre

Subject Co-ordinator

d latesh

Dr.Latesh Malik

Head of Department