DHANALAKSHMI SRINIVASAN ENGINEERING COLLEGE
(AUTONOMOUS)
(Approved by AICTE & Affiliated to Anna University, Chennai)
Accredited with 'A' Grade by NAAC, Accredited by TCS
Accredited by NBA with DME, ECE & EEE
PERAMBALUR - 621 212. Tamil Nadu.
website : www.dsengg.ac.in

U20IT601- FUNDAMENTAL OF DATA SCIENCE

UNIT 1 - INTRODUCTION TO DATA SCIENCE

1.What is data science ?

Data Science combines statistics, maths, specialised programs, artificial intelligence, machine learning etc. Data Science is simply the application of specific principles and analytic techniques to extract information from data used in strategic planning, decision making, etc.

2. Explain the benefits of using statistics in data science.

Statistics help Data scientist to get a better idea of customer's expectation. Using the statistics method Data Scientists can get knowledge regarding consumer interest, behavior, engagement, retention, etc. It also helps you to build powerful data models to validate certain inferences and predictions

3. What are the needs of data science?

The basic needs of data science are:

- Better Decision Making
- Predictive Analysis
- Pattern Discovery

4. List out the various field, where data science are used?

Data Science are used almost everywhere in both commercial and non-commercial settings.

Some of the fields where data science used are

- Healthcare industry
- Retailers
- Financial sectors
- Transportation
- Government sectors
- Universities

5. What are the three sub-phases of data preparations?

The three sub-Phase of data preparations:

•Data cleaning

•Data Integration.

•Data Transformation

¬Graph-based

¬Audio, video and images

¬Streaming

6. What is data cleaning?

Removing missing values, false and inconsistencies across data source.

7. Define Streaming data.

Data that is generated continuously by thousands of data sources, which typically send in the data records simultaneously and in small sizes. Examples are the "What's trending" on Twitter, live sporting or music events, and the stock market.

8. What is Pareto charts?

•It is a graph that indicates the frequency of defects, as well as cumulative impact.

•Pareto charts are useful to find the defects to prioritize in order to observe the greatest Overall improvement.

•It is a combination of a bar graph and line graph.

9. What is recommender system?

Recommender systems are a subclass of information filtering systems, used to predict how ser would rate or score particular objects (movies, music, merchandise, etc.). Recommender systems filter large volume of information based on the data provided by a user and other factors. Recommender systems utilizes algorithms that optimize the analysis of the data to build the recommendations

10. What are various forms of data used in data science?

•The main categories of data are:

¬Structured data

¬Unstructured data

¬ Natural language

¬Machine-generated

11. Explain how a recommender system works.

A recommender system is a system that many consumer-facing, content-driven, online platforms employs to generate recommendations for users from a library of available content. These system generates recommendations based on what they know about the user's tastes from their activities on the platforms.

12. List out the steps involved in data science process.

1. Setting the research goal
2. Retrieving data
3. Data Preparation
4. Data Exploration
5. Data Mining
6. Presentation and automation

13. Mention the input that covers inside the project charter.

- A clear research goal
- The project mission and context
- How you're going to perform your analysis
- What resources you expect to use
- Proof that it's an achievable project, or proof of concepts
- Deliverables and a measure of success
- A timeline

14. List out the various open data sites providers.

Open data site	Description
Data.gov	The home of the US Government's open data
https://open-data.europa.eu/	The home of the European Commission's open data
Freebase.org	An open database that retrieves its information from sites like Wikipedia, Music Brains, and the SEC archive
Data.worldbank.org	Open data initiative from the World Ban

15.Define the techniques to handle missing data.

•Omit the values

•Set value to null

•Impute a static value such as 0 or the mean

•Impute a value from an estimated or theoretical distribution.

•Modeling the value (nondependent)

16. What is outliers?

An outlier is an observation that seems to be distant from other observations or, more specifically, one observation that follows a different logic or generative process than the other observations.The easiest way to find outliers is to use a plot or a table with the minimum and maximum values.

17.What are the different types of Recommender systems?

The two types of recommender systems are

Collaborative filtering – Collaborative filtering is a method of making automatic predictions by using the recommendations of other People.

Content-Based filtering – It is based on the description of an item and a user's choice. As he name suggests, it uses content (keywords) to describe the items, and the ser profile is built to state the type of item this user likes.

18. What are the steps involved in model building?

The model building consists of the following steps such as

a. Selection of a modeling technique and variables to enter in the model

b. Execution of the model

c. Diagnosis and model comparison.

19. What are the various operation involved in combining data.

Two operations to combine information from different data sets.

•The first operation is joining: enriching an observation from one table with information from another table.

•The second operation is appending or stacking: adding the observation of one table to those of another table.

20.What is the difference between a bar graph and a histogram.

Bar chart and Histograms can be used to compare the sizes of the different groups. A bar chart is made up of bars plotted on a chart. A histogram is a graph that represents a frequency

distribution, the height of the bars represent observed frequencies.

UNIT II DATA PREPROCESSING

1. Define Data Preparation.

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis. It is an important step prior to processing and often involves reformatting data, making Corrections to data and the combining of data sets to enrich data.

2. Define Box plots.

•Box plots are a standardized way of displaying the distribution of data based on a five number summary ("minimum", first quartile(Q1), median, third quartile (Q3) and "maximum").

•Median : Middle value of a data set

•First quartile: the middle number between the smallest number and the median.

•Third quartile: the middle nu

3. Define Scatterplots.

A scatterplot is a graph containing a cluster of dots that represents all pairs of scores.

With a little training, you can use any dot cluster as a preview of a fully measured relationship.

4.What is Data integration?

Data integration is the process of combining data from multiple sources into a cohesive and consistent view. This process involves identifying and accessing the different data sources, mapping the data to a common format, and reconciling any inconsistencies or discrepancies between the sources.

5. What is data reduction?

Data reduction is the process of reducing the amount of capacity required to store data.

Data reduction can increase storage efficiency and reduce costs. Storage vendors will often describe storage capacity in terms of raw capacity and effective capacity, which refers to data after the reduction.

6. What is data transformation?

Data transformation is the process of converting data from one format to another, typically from the format of a source system into the required format of a destination system.

7.Define Data discretization

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss.

8. What Is Logistic Loss?

Logistic Loss (logloss) is another evaluation method that is commonly used in classification problems. The basic idea is to try to measure the likelihood of the similarity between the predictive values (probabilities) and the actual values

9.Define AUC

(Area Under ROC Curve) which is simply used as a single value to be compared between different models for evaluating their performance.

10.What are the four steps of exploratory data analysis?

Steps Involved in Exploratory Data Analysis

- Data Collection. Data collection is an essential part of exploratory data analysis. ...
- Data Cleaning. Data cleaning refers to the process of removing unwanted variables and values from your dataset and getting rid of any irregularities in it. ...
- Univariate Analysis. ...
- Bivariate Analysis.

11. What is EDA?

In data mining, Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics, often with visual methods. EDA is used for seeing what the data can tell us before the modeling task.

12.What are the components of EDA?

The main components of the EDA pattern are the following:

- Event specifications.
- Event processing.
- Event tooling.
- Enterprise integration.
- Sources and targets.

13.What are the different types of EDA?

The four types of EDA are univariate non-graphical, multivariate non- graphical, univariate graphical, and multivariate graphical.

14.What are the underlying principles of exploratory data analysis?

The main underlying principles of an EDA are-

- The aim should be to uncover information that should lead to showing patterns and trends.
- Missing values and outliers need to be given proper consideration

- The relationship between different variables must be established.
- A suitable technique of variate analysis should be chosen for the target to be achieved.

15.Define ROC curve

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

16.What is the philosophy of EDA?

EDA is a method or philosophy that aims to uncover the most important and frequently overlooked patterns in a data set. We examine the data and attempt to formulate a hypothesis. Statisticians use it to get a bird eyes view of data and try to make sense of it.

17.What is the purpose of EDA in data science?

Exploratory Data Analysis (EDA) is widely used by Data Scientists while analyzing and investigating Data sets, summarizing the main characteristics of data to the visualizing method. It helps the Data Scientist to discover Data Patterns, Spot anomalies. hypothesis testing, and or assumption.

UNIT-III BASIC MACHINE LEARNING ALGORITHMS

1.What are basic machine learning algorithms?

Algorithm Basics. The word Algorithm means " A set of finite rules or instructions to be followed in calculations or other problem-solving operations " Or " A procedure for solving a mathematical problem in a finite number of steps that frequently involves recursive operations".

2.What is supervised learning?

It is used for the structured dataset. It analyzes the training data and generates a function that will be used for other datasets.

3.What is Unsupervised learning?

It is used for raw datasets. Its main task is to convert raw data to structured data. In today's world, there is a huge amount of raw data in every field. Even the computer generates log files which are in the form of raw data. Therefore it's the most important part of machine learning.

4. Define Decision Tree

The decision tree is a widely utilized supervised machine learning algorithm that finds applications in tasks such as classification and regression. It adopts a tree-like structure, wherein

internal nodes symbolize features or attributes, branches represent decision rules based on those attributes, and leaf nodes correspond to outcomes or predictions.

5. What is Linear Regression?

It is the most well-known and popular algorithm in machine learning and statistics. This model will assume a linear relationship between the input and the output variable. It is represented in the form of a linear equation which has a set of inputs and a predictive output. Then it will estimate the values of the coefficient used in the representation.

6. Define k-Nearest Neighbors (k-NN)

This algorithm is used for classification problems and statistical problems as well.

Its model is to store the complete dataset. By using this algorithm, prediction is done by searching the entire training data for k instances.

7. What is Naive Bayes algorithm in machine learning?

The Naïve Bayes classifier is a supervised machine learning algorithm, which is used for classification tasks, like text classification. It is also part of a family of generative learning algorithms, meaning that it seeks to model the distribution of inputs of a given class or category.

8. How does the Naive Bayes classification algorithm work?

The Naive Bayes classification algorithm is a probabilistic classifier. It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive.

9. Is naive Bayes a clustering algorithm?

Naive Bayes inference is a very common technique for performing data classification, but it's not generally known that Naive Bayes can also be used for data clustering.

10. What is the difference between Bayes and naive Bayes?

Bayes theorem provides a way to calculate the conditional probability of an event based on prior knowledge of related conditions. The naive Bayes algorithm, on the other hand, is a machine learning algorithm that is based on Bayes' theorem, which is used for classification problems.

11. What is random forest in machine learning advantages and disadvantages?

Advantages and Disadvantages of Random Forest

It requires much computational power as well as resources as it builds numerous trees to combine their outputs. It also requires much time for training as it combines a lot of decision trees to determine the class.

12. What is a random forest in machine learning?

Random forest is a commonly-used machine learning algorithm trademarked by Leo Breiman and Adele Cutler, which combines the output of multiple decision trees to reach a single result

13. Why do we use random forest in machine learning?

A. Random Forest is a popular machine learning algorithm used for classification and regression tasks due to its high accuracy, robustness, feature importance, versatility, and scalability. Random Forest reduces overfitting by averaging multiple decision trees and is less sensitive to noise and outliers in the data.

14. What is feature generation and feature selection?

Feature Generation (also known as feature construction, feature extraction or feature engineering) is the process of transforming features into new features that better relate to the target.

15. What are the three types of feature selection?

There are three types of feature selection: Wrapper methods (forward, backward, and stepwise selection), Filter methods (ANOVA, Pearson correlation, variance thresholding), and Embedded methods (Lasso, Ridge, Decision Tree)

UNIT-IV CLUSTERING

1. What is Clustering?

Clustering is a type of unsupervised learning method of machine learning. In the unsupervised learning method, the inferences are drawn from the data sets which do not contain labeled output variable.

2. Define DBSCAN

DBSCAN groups data points together based on the distance metric. It follows the criterion for a minimum number of data points. It can discover clusters of different shapes and sizes from a large amount of data, which is containing noise and outliers

3. Which of the following is required by K-means clustering?

K-means clustering follows partitioning approach.

4. Which of the following is true about k-means clustering?

Because it employs the mean of cluster pieces of data to locate the cluster center, the K-Means clustering technique is very sensitive to outliers

5. What is finally produced by hierarchical clustering?

Explanation: Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram.

6.Which of the following clustering is hierarchical clustering?

Agglomerative Hierarchical clustering Technique: In this technique, initially each data point is considered as an individual cluster. At each iteration, the similar clusters merge with other clusters until one cluster or K clusters are formed. The basic algorithm of Agglomerative is straight forward.

7. What does hierarchical clustering show?

Hierarchical clustering is separating data into groups based on some measure of similarity, finding a way to measure how they're alike and different, and further narrowing down the data.

8. Which of the following function is used for k-means clustering?

The Elbow method is the best way to find the number of clusters. The elbow method constitutes running K-Means clustering on the dataset. Next, we use within-sum-of-squares as a measure to find the optimum number of clusters that can be formed for a given data set.

9. Is k-means deterministic or non-deterministic?

The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids.

10. What is the function used for k-means?

K-means algorithm requires users to specify the number of cluster to generate. The R function kmeans() [stats package] can be used to compute k-means algorithm. The simplified format is kmeans(x, centers), where "x" is the data and centers is the number of clusters to be produced.

11. What is the function for clustering?

Clustering is the task of dividing the unlabeled data or data points into different clusters such that similar data points fall in the same cluster than those which differ from the others. In simple words, the aim of the clustering process is to segregate groups with similar traits and assign them into clusters.

12. Is one of the most commonly used methods for clustering is the K-Means algorithm?

K-means clustering is a widely used method for cluster analysis where the aim is to partition a set of objects into K clusters in such a way that the sum of the squared distances between the objects and their assigned cluster mean is minimized.

UNIT –V - DATA VISUALIZATION

1. What is data visualization.

Graphical way of representing the dataset is known as Data visualization.

2. Write a comment to import matplotlib.

import matplotlib as mpl

import matplotlib.pyplot as pl

3. What are the two interface of Matplotlib.

Feature of Matplotlib is its dual interfaces: a convenient MATLAB-style state-based interface, and a more powerful object-oriented interface.

4. Define Scatter plots.

Scatter plots is similar to line plot. Instead of points being joined by line segments, here the points are represented individually with a dot, circle, or other shape.

5. What is seaborn?

Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

9.Mention the functionality of Seaborn?

Seaborn offers the following functionalities:

1. Dataset oriented API to determine the relationship between variables.
2. Automatic estimation and plotting of linear regression plots.
3. It supports high-level abstractions for multi-plot grids.
4. Visualizing univariate and bivariate distribution.

10. What is the use of Statsmodels in Python?

Python StatsModels allows users to explore data, perform statistical tests and estimate statistical models. It is supposed to complement to SciPy's stats module. It is part of the Python scientific stack that deals with data science, statistics and data analysis.

11. Mention the uses of Density Maps

Density mapping is simply a way to show where points or lines may be concentrated in a given area.Using Density Maps with Python

Here, we will be using a worldwide dataset of earthquakes and their magnitudes.

12.Where is ploty used?

Plotly can also be used to style interactive graphs with Jupyter notebook. Figure

Converters which convert matplotlib, ggplot2, and IGOR Pro graphs into interactive, online Graphs

13 What is binning?

Data binning is a type of data preprocessing, a mechanism which includes also dealing with missing values, formatting, normalization and standardization. Binning can be applied to convert numeric values to categorical or to sample numeric values

14.What are the features of Matplotlib

Matplotlib supports all the popular charts (lots, histograms, power spectra, bar charts, error charts, scatterplots, etc.) right out of the box. There are also extensions that you can use to create advanced visualizations like 3-Dimensional plots, etc.

15.What is a Contour plot?

A contour plot is a graphical technique which portrays a 3-dimensional surface in two dimensions. Such a plot contains contour lines, which are constant z slices. To draw the contour line for a certain z value, we connect all the (x, y) pairs, which produce the value z.

16.Define Density plot in Python.

A 2D histogram contour plot, also known as a density contour plot, is a 2-dimensional generalization of a histogram which resembles a contour plot but is computed by grouping a set of points specified by their x and y coordinates into bins, and applying an aggregation function such as count or sum (if z is provided) to compute the value to be used to compute contours.

17. What are the main types of visualizations?

Main Types of Visualizations

- Sketch.
- Diagram.
- Blueprint.
- Landscape.
- Artist Impression.
- Heat map.
- Roadmap.
- Scenario.

18. What are data visualization methods?

Data visualization is the graphical representation of information and data. By using visual elements like **charts, graphs, and maps**, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

19. What are the key components of data visualization?

Data visualization components

- Bar charts.
- Line charts.
- Area charts.
- Pie charts.
- Scatter charts.
- Bubble charts.

20. What are some common tools used for data visualization?

The best data visualization tools include Google Charts, Tableau, Grafana, Chartist. js, FusionCharts, Datawrapper, Infogram, ChartBlocks, and D3. js. The best tools offer a variety of visualization styles, are easy to use, and can handle large data sets.

DHANALAKSHMI SRINIVASAN ENGINEERING COLLEGE
(AUTONOMOUS)
(Approved by AICTE & Affiliated to Anna University, Chennai)
Accredited with 'A' Grade by NAAC, Accredited by TCS
Accredited by NBA with BME, ECE & EEE
PERAMBALUR - 621 212. Tamil Nadu.
website : www.dsengg.ac.in

U20IT601- FUNDAMENTAL OF DATA SCIENCE
UNIT I - INTRODUCTION TO DATA SCIENCE
PART-B

1. How to Install the R-studio?
2. Briefly describe the use of statistics in Data Science.
3. Explain different stages of data Science?
4. How to getting the data in and out of R-Language?
5. Analyze the descriptive statistics
6. Explain general techniques for handling volumes of data?
7. Discuss in detail about representing matrices by decompositions (SVD & PCA)
8. Write short notes.on R/python
9. Explain R Objects?
10. Describe the main features of a big data in detail.
11. What is matrices? Describe matrices to represent relationship between data
12. Explain current landscape of perspectives in data science
13. Explain in detail about statistical Interference
14. Write short notes on hypothesis testing

UNIT II - DATA PREPROCESSING

1. Discuss in detail about basic tools of EDA
2. Explain in detail about evaluation of classification methods
3. Write short notes on data pre-processing
4. Explain the tools of EDA
5. Write short notes on philosophy of EDA
6. Write and explain the following i) Confusion matrix ii) Students T-tests and ROC curves

UNIT III-MACHINE LEARNING ALGORITHMS

1. Describe the role of correctness in machine learning.
2. Illustrate the curse of dimensionality in detail.

3. Explain the goodness of fitting in multiple regression model
4. Describe regularization in detail.
5. Explain the simple linear regression model in detail.
6. Discuss random forests in detail.
7. Write Python program to create a decision tree.
8. Discuss decision tree in detail.
9. Describe the naïve bayes ensemble methods
10. Discuss in detail about basic machine learning algorithms
11. Explain in detail about random forests
12. Write short notes,on linear regression and logistic regression

UNIT IV-CLUSTERING

1. Develop the Association rule mining
2. Illustrate K Means Clustering Algorithm
3. Analyze in detail about DBSCAN
4. Explain in detail about different clustering approaches
5. Explain in detail about hierarchical agglomerative clustering
6. Examine K-means lolyds algorithm
7. Illustrate the K-NN Classifier
8. Illustrate and solve the problem of given values s1=(5,8,3,4,3,6,6,5) and s2=(7,4,34,7,7,1,5) using DBSCAN algorithm .
9. Illustrate and solve the K Means Clustering Algorithm of data sets in to two clusters X=(1,2,2,3,45),Y=(1,1,3,2,3,5)

UNIT V-DATA VISUALIZATION

1. Discuss briefly about the types of data visualization.
2. Explain and identify the tools used in data visualization.
3. Explain in detail about the principles of data visualization