CSE-Data Science/Data Science, IV semester

CD404 INTRODUCTION TO DATA SCIENCE

Unit – I: Introduction

Introduction to Data Science – Evolution of Data Science – Data Science Roles – Stages in a Data Science Project – Applications of Data Science in various fields – Data Security Issues.

Unit – II: Data Collection and Data Pre-Processing

Data Collection Strategies – Data Pre-Processing Overview – Data Cleaning – Data Integration and Transformation – Data Reduction – Data Discretization.

Unit – III: Exploratory Data Analytics

Descriptive Statistics – Mean, Standard Deviation, Skewness and Kurtosis – Box Plots – Pivot Table – Heat Map – Correlation Statistics – ANOVA.

Unit – IV: Model Development

Simple and Multiple Regression – Model Evaluation using Visualization – Residual Plot – Distribution Plot – Polynomial Regression and Pipelines – Measures for In-sample Evaluation – Prediction and Decision Making.

Unit – V: Model Evaluation

Generalization Error – Out-of-Sample Evaluation Metrics – Cross Validation – Overfitting – Under Fitting and Model Selection – Prediction by using Ridge Regression – Testing Multiple Parameters by using Grid Search.

REFERENCES:

1. JojoMoolayil, “Smarter Decisions : The Intersection of IoT and Data Science”,PACKT, 2016.

2. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

3. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

4. Raj, Pethuru, “Handbook of Research on Cloud Infrastructures for Big DataAnalytics”, IGI Global.

List of Experiments:

1. READING AND WRITING DIFFERENT TYPES OF DATASETS using Python

   a. Reading different types of data sets (.txt, .csv) from web and disk and writing in file in specific disk location.
   b. Reading Excel data sheet in python.
   c. Reading XML dataset in python.
2. VISUALIZATIONS:
   a. Find the data distributions using box and scatter plot.
   b. Find the outliers using plot.
   c. Plot the histogram, bar chart and pie chart on sample data
3. EXPLORATORY DATA ANALYSIS (EDA): Perform EDA on Credit Card Fraud Detection Dataset (open source dataset) for analyzing the data.
4. LINEAR REGRESSION MODEL FOR PREDICTION: Apply Regression Model techniques to predict the future values of data on the open source available datasets.
5. LOGISTIC REGRESSION MODEL: Import the Red-Wine dataset from the UCI Machine Learning Repository having three qualities of wines. Apply logistic regression model for multi-class classification of the wine categories.
6. MODEL EVALUATION USING RESIDUAL PLOT: Plotting Accuracy and Error Metrics against number of iterations for evaluation of model performance.
7. EVALUATING UNDER-FITTING AND OVER-FITTING: Plotting Learning curves for model evaluation for Under-fitting and Over-fitting

CD404: Introduction to Data Science:

Unit I: Introduction to Data Science

Introduction to Data Science:

- Data Science is an interdisciplinary field that utilizes scientific methods, algorithms, and systems to extract insights and knowledge from data.

- It involves various processes including data collection, preparation, analysis, visualization, and interpretation.

- The field has emerged due to the exponential growth of data and advancements in technology.

Evolution of Data Science:

- Data Science has evolved from traditional statistics and computer science disciplines.

- It has roots in statistics, machine learning, data mining, and big data technologies.

- The rise of big data and the need for extracting actionable insights have driven the growth of Data Science.

Data Science Roles:

- Data Scientist: Analyzes complex datasets to derive insights and build predictive models.

- Data Engineer: Develops and manages data infrastructure and systems.

- Data Analyst: Examines data to identify trends, patterns, and insights.

- Domain Expert: Provides subject matter expertise to interpret data in specific domains.

Stages in a Data Science Project:

1. Problem Definition: Identifying the problem statement and defining project objectives.

2. Data Collection: Gathering relevant data from various sources.

3. Data Preprocessing: Cleaning, integrating, transforming, and preparing the data for analysis.

4. Exploratory Data Analysis (EDA): Understanding the characteristics and patterns in the data.

5. Model Development: Building predictive or descriptive models using machine learning or statistical techniques.

6. Model Evaluation: Assessing the performance of models and fine-tuning parameters.

7. Deployment and Interpretation: Implementing the model in real-world scenarios and interpreting results for decision-making.

Applications of Data Science:

- Healthcare: Predictive analytics for disease diagnosis and treatment optimization.

- Finance: Risk assessment, fraud detection, algorithmic trading.

- Marketing: Customer segmentation, personalized recommendations, churn prediction.

- E-commerce: Product recommendation systems, market basket analysis.

- Transportation: Route optimization, demand forecasting, traffic management.

Data Security Issues:

- Data Privacy: Protecting sensitive information and ensuring compliance with regulations.

- Data Breaches: Preventing unauthorized access to data and securing networks and systems.

- Ethical Considerations: Addressing ethical dilemmas related to data collection, usage, and sharing.

- Cybersecurity: Implementing measures to safeguard data against cyber threats and attacks.

Unit II: Data Collection and Data Pre-Processing

Data Collection Strategies:

- Surveys and Questionnaires

- Observational Studies

- Experiments

- Web Scraping

- APIs

Data Pre-Processing Overview:

- Data Cleaning
- Data Integration
- Data Transformation
- Data Reduction
- Data Discretization

Data Cleaning:

- Handling Missing Values
- Outlier Detection
- Data Validation
- Standardization and Normalization

Data Integration and Transformation:

- Entity Resolution
- Schema Integration
- Feature Engineering
- Encoding Categorical Variables

Data Reduction:

- Dimensionality Reduction
- Sampling Methods

Data Discretization:

- Binning
- Equal-Width and Equal-Frequency Binning

- Discretization Techniques

Unit III: Exploratory Data Analytics

Descriptive Statistics:

- Mean: Average value of a dataset.
- Standard Deviation: Measure of the dispersion of values around the mean.
- Skewness and Kurtosis: Measures of the asymmetry and peakedness of a distribution.
- Box Plots: Visual representation of the distribution of data.
- Pivot Table: Tool for summarizing and analyzing data in spreadsheet programs.
- Heat Map: Visual representation of data where values are represented as colors.

Correlation Statistics:

- Measures the strength and direction of the relationship between two variables.
- Pearson correlation coefficient: Measures linear correlation between variables.
- Spearman rank correlation coefficient: Measures monotonic relationship between variables.
- Kendall's tau: Measures correlation for ordinal data.

ANOVA (Analysis of Variance):

- Statistical technique used to analyze the differences between group means in a sample.
- Determines whether there are statistically significant differences between the means of three or more groups.

Unit IV: Model Development

Simple and Multiple Regression:

- Simple Regression: Predicting a dependent variable based on one independent variable.

- Multiple Regression: Predicting a dependent variable based on multiple independent variables.

Model Evaluation using Visualization:

- Residual Plot: Plot of the residuals (difference between observed and predicted values) against the independent variable.

- Distribution Plot: Visualization of the distribution of data points.

Polynomial Regression and Pipelines:

- Polynomial Regression: Extends linear regression to fit nonlinear relationships.

- Pipelines: Sequence of data processing components that are chained together to process data.

Measures for In-sample Evaluation:

- R-squared (Coefficient of Determination): Proportion of the variance in the dependent variable that is predictable from the independent variable(s).

- Mean Squared Error (MSE): Average of the squares of the errors.

Prediction and Decision Making:

- Using models to make predictions based on input data.

- Decision making involves interpreting model outputs and taking appropriate actions based on predictions.

Unit V: Model Evaluation

Generalization Error:

- Error rate on new, unseen data.

- Indicates how well the model performs on data it hasn't seen during training.

Out-of-Sample Evaluation Metrics:

- Evaluate model performance on a separate test dataset.

- Metrics include accuracy, precision, recall, F1-score, ROC curves, etc.

Cross Validation:

- Technique for assessing how the results of a statistical analysis will generalize to an independent dataset.

- Common methods include k-fold cross-validation and leave-one-out cross-validation.

Overfitting and Underfitting:

- Overfitting: Model learns to capture noise in the training data and performs poorly on new data.

- Underfitting: Model is too simple to capture the underlying structure of the data.

Model Selection:

- Choosing the best model among different candidate models based on evaluation metrics.

- Involves comparing performance on validation datasets and selecting the model with the best generalization error.

Prediction using Ridge Regression:

- Ridge Regression: Technique for mitigating multicollinearity in multiple regression models by adding a penalty term to the loss function.

Testing Multiple Parameters using Grid Search:

- Grid Search: Exhaustive search over specified parameter values for an estimator.

- Helps find the optimal parameters for a model.