FOUNDATION OF DATA SCIENCE

ENCT 202

Lecture : 3 **Year : II**
Tutorial : 1 **Part : I**
Practical : 3

Course Objectives:
The objective of this course is to introduce the core concepts, tools, and methodologies of data science, focusing on the tools and techniques needed to analyze and interpret data effectively. Using data science tools, students will cover the entire data science process, from data acquisition, data manipulation, visualization, probability, statistics, and machine learning, with applications in business and engineering.

1 Introduction to Data Science (3 hours)

1.1 Overview of data science
1.2 Jargons of data science
1.3 Modern data ecosystem
1.4 Data science lifecycle
1.5 Trends, markets and applications of data science
1.6 Tools and technologies in data science
1.7 Data scientist and their roles

2 Mathematics for Data Science (10 hours)

2.1 Introduction to linear algebra for data science
2.2 Vectors, matrices and matrix factorization
2.3 Gradient descent for optimization
2.4 Introduction to probability and random variable
2.5 Probability distributions: Normal, Bernoulli, Binomial, Poisson
2.6 Descriptive and inferential statistics
2.7 Central limit theorem and sample distribution concepts
2.8 Normal approximation; hypothesis testing procedures: Tests about the mean of a normal population
2.9 The t-test, Z-tests for differences between two populations means, the two-sample t-test, confidence interval for mean of normal population
2.10 ANOVA

3 Data Understanding and Preprocessing (10 hours)

3.1 Types of data: Structured, unstructured, semi-structured
3.2 Data preprocessing requirements
3.3 Data sources and collection methods

3.4 Data cleaning and preparation
3.5 Data wrangling and associated tools
3.6 Data enrichment, validation and publishing
3.7 Data transformation and normalization
3.8 Dimensionality reduction linear factor model, principal component analysis (PCA)

4 Data Analysis (8 hours)

4.1 Data analytics: Descriptive, diagnostic, predictive and prescriptive analytics
4.2 Exploratory data analysis using descriptive statistics
4.3 Data visualization
4.4 Data visualization techniques
4.5 Principles of effective data visualization
4.6 Feature engineering and other aspects of data manipulation

5 Regression and Predictive Modeling (5 hours)

5.1 Empirical models, simple linear regression, MLE and least square estimator
5.2 Multiple linear regression, matrix approach to multiple linear regression, polynomial regression models, categorical regressors, indicator variables, selection of variables and model building
5.3 Logistic regression

6 Modeling and Validation Processes (6 hours)

6.1 Introduction to machine learning
6.2 Introduction to supervised, unsupervised and reinforcement learning
6.3 Modeling process, training /validating model, cross validation methods, predicting new observations interpretation
6.4 Measures for model performance and evaluation: Classification accuracy, confusion matrix, sensitivity, specificity, precision, recall, F-score, ROC curve, clustering performance measures, other measures

7 Ethics and Recent Trends (3 hours)

7.1 Ethical considerations in data science
7.2 Data privacy regulations
7.3 Responsible data usage
7.4 The five Cs
7.5 Future trends

Tutorial (15 hours)

1. Solution of data problems using linear algebra, vectors and matrices
2. Solution of the problems related probability and statistics to understand application in data science

3. Identification of the data types and performing data cleaning, transformation, wrangling, and dimensionality reduction Including EDA and feature engineering
4. Solution of the problem related to linear and logistic regression
5. Understanding machine learning basics by model training, cross-validation, and performance evaluation

Practical **(45 hours)**

1. Get acquainted with data science tools and perform statistical analysis
2. Hypothesis tests (e.g., t-tests, Z-tests) on sample datasets to compare population means
3. Simulate and apply the central limit theorem (CLT) to demonstrate how sample distributions converge to a normal distribution
4. Perform data wrangling and ETL processes on a dataset, followed by exploratory data analysis (EDA)
5. Utilize tools to create effective data visualizations (e.g., line charts, bar charts, heat maps, box plots) to derive key insights from the dataset
6. Implement feature extraction and selection techniques, including experimenting with encoding methods like one-hot encoding and creating new features based on domain expertise
7. Develop a simple linear regression model, extend it to multiple linear regression with several variables, and visualize both the regression line and residual plots
8. Apply logistic regression and evaluate the model using metrics such as accuracy, precision, recall, and the ROC curve
9. Apply K-means clustering and assess cluster quality using evaluation metrics like the silhouette score

By the end of the practical, students are required to submit a project where they develop a prototype to solve a real-world problem.

Final Exam

The questions will cover all the chapters in the syllabus. The evaluation scheme will be as indicated in the table below:

Chapter	Hours	Marks distribution*
1	3	6
2	10	12
3	10	12
4	8	9
5	5	6
6	6	9
7	3	6
Total	**45**	**60**

* There may be minor deviation in marks distribution.

References

1. Ozdemir, S. (2016). Principles of Data Science. Germany: Packt Publishing.
2. Maheshwari A. (2018). Data Science for Dummies, Wiley.
3. Grus, J. (2019). Data Science from Scratch: First Principles with Python. United States: O'Reilly Media.
4. Bruce, P., Bruce, A. (2017). Practical Statistics for Data Scientists: 50 Essential Concepts. United States: O'Reilly Media.
5. VanderPlas, J. (2016). Python Data Science Handbook: Essential Tools for Working with Data. United States: O'Reilly Media.
6. Provost, F., Fawcett, T. (2013). Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking. United States: O'Reilly Media.