SEMESTER: I **HOURS/WEEK : 5**
Code: P21DS101 **CREDITS : 4**

Core I: Mathematical Foundation for Data Science

Course Objectives

Every day all over the world, large amounts of data are generated by various businesses, research organizations and social media. Data Science is the study, application, and development of methods to learn from this data. So that many improvements can be made in products, services, advertising campaigns, public health and safety and others. Linear algebra, geometry, calculus and probability play a fundamental role in the theory of Data Science. This course introduces the basic notions of vector spaces, inner product, matrix decomposition, derivatives and probability distributions.

Course Outcomes

On successful completion of this course, students will be able to

CO1: Solve systems of linear equations by use of the matrix
CO2: Determine the Orthogonality and Basis
CO3: Compute and use eigenvectors and eigenvalues
CO4: Analyze the gradients and PDE
CO5: Use the probability concepts in Data science

Unit - I: Linear Algebra

Systems of Linear Equations – Matrices - Solving Systems of Linear Equations - Vector Spaces - Linear Independence - Basis and Rank - Linear Mappings - Affine Spaces.

Unit - II: Analytic Geometry

Norms - Inner Products - Lengths and Distances - Angles and Orthogonality - Orthonormal Basis - Orthogonal Complement - Inner Product of Functions - Orthogonal Projections - Rotations

Unit - III : Matrix Decompositions

Determinant and Trace - Eigenvalues and Eigenvectors - Cholesky Decomposition – Eigen decomposition and Diagonalization - Singular Value Decomposition Matrix Approximation - Matrix Phylogeny -

Unit – IV : Vector Calculus

Differentiation of Univariate Functions - Partial Differentiation and Gradients - Gradients of Vector-Valued Functions - Gradients of Matrices - Useful Identities for Computing Gradients - Backpropagation and Automatic Differentiation - Higher-Order Derivatives - Linearization and Multivariate Taylor Series

Unit – V : Probability and Distributions

Probability and Distributions: Construction of a Probability Space - Discrete and Continuous Probabilities - Sum Rule, Product Rule, and Bayes' Theorem - Summary Statistics and Independence - Gaussian Distribution.

Text Book

1. Marc Peter Deisenroth, A. Aldo Faisal, Cheng Soon Ong, " *Mathematics for Machine Learning*", Cambridge Press, 2019 (Chapters 2, 3, 4,5,6)

References

1. Gilbert Strang,*"Introduction to Linear Algebra"*, 3ed, Cambridge Press, 2003.
2. M. D. Weir, J. Hass, and G. B. Thomas, *"Thomas' calculus*", Pearson, 2016.

SEMESTER: I **HOURS/WEEK : 5**
Code: P21DS102 **CREDITS : 4**

Core II: Problem Solving Using Python and R

Course Objectives

This course introduces students the language features of both Python and R. Specifically, data structures, regular expressions, data visualization and internet programming features are introduced.

Course Outcomes

On successful completion of this course, students will be able to:
CO1: Develop applications using Python data structures
CO2: Develop object oriented programs in Python
CO3: Manipulate files using Python
CO4: Access internet and database data
CO5: Develop R programs for data visualization

Unit-1. Python Basics, Functions, Loops and Strings

Variables – Getting Inputs – Conditions – Catching exceptions – Function calls – Built-in functions – Type conversion functions and math functions – Parameters and arguments –While statement – Infinite loops – Continue statement – For loops – Strings – Slice –– The in operator – String comparison – String methods – parsing strings – Format operator.

Unit-2. Files and Data Structures

Opening files – Text files – Reading files – Searching through files – Writing files – Traversing list – List operations – List slice – List methods – Deleting elements – Built-in list functions – Objects, value and aliasing – List arguments. Dictionaries – Files and dictionaries – Looping and dictionaries – Tuples – Comparing tuples – Tuple assignments – Dictionaries and tuples – Tuples as keys in dictionaries

Unit-3. Object Oriented Programming and Internet Programming

Creating objects – Encapsulation – Classes as types – Object lifecycle – Instances – Inheritance. Regular expressions – Character matching – Extracting data – Escape character – Designing simple web browser using sockets – Retrieving images using HTTP – Retrieving web pages using urllib – Reading binary files using urllib – Accessing data from databases

Unit-4. Functional Programming with R

Variables - Vector, matrix, arrays – List – Data Frames – Functions – Strings – Factors – Loops – Packages –Date and Time – Files – Make packages

Unit-5. Data Analysis using R

Data analysis using R - Working with data frames - R inbuilt data sets - Visualisation using ggplot2 - Creating documentation and reports - Creating simple dashboards using shiny

Text Books

1. Allen B. Downey, ―Think Python: How to Think like a Computer Scientist, 2ndedition, Updated for Python 3, O_Reilly Publishers, 2016
2. Charles R. Severance, Python for Everybody: *"Exploring data using Python 3"*,Schroff Publishers, 1ed, 2017, ISBN 978-9352136278.
3. Richard Cotton, "Learning R", O'Reilly, 2013

References

1. Zed Shaw's , Learn Python the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code, Addison-Wesley Professional; 3 edition, 2013
2. Robert Sedgewick, Kevin Wayne, Robert Dondero, Introduction to Programming in Python: An Inter -disciplinary Approach, Pearson India Education Services Pvt. Ltd., 2016.
3. Wesley J Chun, Core Python Programming , 2nd edition, Prentice Hall ,2009
4. Colin Gillespie, Robin Lovelace, and EfficientR Programming: A Practical Guide to Smarter Programming, "O'Reilly Media, Inc.", 2016
5. Paul Teetor, R Cookbook-Proven Recipes for Data Analysis, Statistics, and Graphics, O'Reilly Media, 2011

SEMESTER: I **HOURS/WEEK : 5**
Code: P21DS103 **CREDITS : 4**

Core III: NoSQL Database Management

Course Objectives

The widespread emergence of big data storage needs has driven the development and adoption of a new class of non-relational databases commonly referred to as NoSQL databases. This course will explore the origins of NoSQL databases and the characteristics that distinguish them from traditional relational database management systems. Core concepts of NoSQL databases will be presented, followed by an exploration of how different database technologies implement these core concepts.

Course Outcomes

On successful completion of this course, students will be able to:

CO1: Model data using ER diagrams
CO2: Demonstrate competency in designing SQL and NoSQL database management systems.
CO3: Demonstrate competency in describing how NoSQL databases differ from relational databases
CO4: Demonstrate competency in selecting a particular NoSQL database for specific use cases.
CO5: Implement databases using SQL, MongoDB and Neo4J

Unit-1. Structured Query Language-I

ER Model: Entity types, Attribute types, Relationship types – Weak entity types, Ternary relationship types – Examples of ER model. Enhanced ER model: Specialization/Generalization – Categorization - Aggregation – Examples of EER. Relational DB Process and outcome approach - Simple Queries on one table – First look at joins – Sub queries.

Unit-2.Structured Query Language-II

Self Joins: Self relationships, Questions involving Both – Multiple relations between tables – Set operations – Aggregate Operations – Window functions – Efficiency considerations: Indexing and Join Techniques.

Unit-3. MongoDB-I

Introduction: MongoDB document, collection and database – Basic Operations – Datatypes – Creating, deleting, updating documents: insert, batch insert, remove, find, findone, update – arrays – insert – Updating multiple documents

Unit-4. MongoDB-II

Comparison operators – OR and NOT queries – Querying arrays – Querying on embedded documents – WHERE queries – Limits, skips and sort – Compound Index – Unique index – Sparse Index – Pipeline aggregation: MATCH, PROJECT, GROUP and UNWIND clauses.

Unit-5. Neo4J and Cypher

Labeled Property Graph Model – Querying graphs using Cypher: CREATE AND ASSERT, MATCH, WHERE and RETURN clauses– ORDER BY – WITH clause – Case Study: Telent.net Social recommendations application.

Text Books

1. Clare Churcher. *Beginning SQL Queries: From Novice to Professional*, APress, 2ed, 2016. ISBN 978-1-4842-1954-6
2. WilfriedLemahieu, SeppevandenBroucke and Bart Baesens. *Principles of Database Management: The Practical Guide to Storing, Managing and Analyzing Big and Small Data*, Cambridge University Press, 2018. ISBN 978-1-107-18612-5 (Chapter 3 ER diagram only)
3. Kristina Chodorow, *MongoDB: The Definitive Guide*, 2ed, Oreilly Publishers
4. Ian Robinson, Jim Webber and Emil Eifrem. Graph Databases: New Opportunities for connected data. 2ed, Oreilly Publishers. ISBN 978-1491930892.

References

1. Eric Redmond; Jim R. Wilson. *Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement*. Pragmatic Bookshelf. 2012. ISBN: 1934356921Pramod J. Sadalage; Martin

Fowler. *NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence*. Addison-Wesley. 2012 ISBN: 0321826620
2. Adam Fowler. *NoSQL for Dummies*. John Wiley. 2015. ISBN 978-1-118-90574-6
3. Guy Harrison. *Next Generation Databases*. APress. 2016. 978-1-484213-30-8
4. Thomas M. Connolly and Carolyn E. Begg. Database Systems: "*A Practical Approach to Design, Implementation, and Management"*, 6th Edition, Pearson, 2015.

SEMESTER: I **HOURS/WEEK : 5**
Code: P21DS1:1 **CREDITS : 4**

Elective-I: Essential Statistics for Data Science

Course Objectives

This course covers topics in Statistics from basics to advanced level that every Data Science student should master and apply for the industry applications. Great depth of coverage for the topics from Regression Analysis is also given in this course.

Course Outcomes

On successful completion of this course, students will be able to
CO1: Identify the methods of descriptive statistics and variablity.
CO2: Examine the different tests of the statistical inferences
CO3: Demonstrate the nonparametric statistics methods
CO4: Classify the different types of regression methods for data analytics
CO5: Analyze the different properties of the regression methods.

Unit I: Descriptive Statistics

Introduction to Statistics - Organizing Data Using Tables and Graphs- Measures of Central Tendency: Mode – Median – Mean. Measures of Variability: Variability – Range - Interquartile Range - Standard Deviation.

Unit II: Inferential Statistics – I

Sampling Distribution of Means: Sampling Distribution - Central Limit Theorem. Hypothesis Testing: Hypothesis Testing Steps -Effect Size for a Z-Test - Assumptions – Errors – Power. One-Sample t Test: t-Statistics – t- Distributions - One-Sample t Test – Effect Size – Assumptions. Two-Sample t Test: Independent Samples Design: Calculations – Hypothesis Testing – Effect Size – Assumptions. Two-Sample t Test: Related Samples Design: Calculations – Hypothesis Testing – Effect Size – Assumptions.

Unit III: Inferential Statistics - II

Confidence Interval versus Point Estimation: Introduction- Point Estimates - Confidence Intervals – One Sample t- Test - Two-Sample t Test: Independent Samples Design – Repeated Measure Design - Degree of Confidence Vs. Degree of Specificity One-Way Analysis of Variance: Introduction – Variance – F- statistics – Hypothesis Testing with F- Statistic - F- Distribution Table - Notations for ANOVA - Calculations – Hypothesis Testing – Effect Size – Assumptions. Chi-Square: Chi-Square - Chi-Square Statistic – Assumptions- Goodness of Fit - Goodness of Fit for Known Proportions- Goodness of Fit for No Preference – Test of Independence - Nonparametric Statistics for Ordinal Data: Mann-Whitney U Test - Kruskal-Wallis H Test. Correlation: Introduction – Scatter Plot - Pearson Product Moment Correlation - Hypothesis Testing - Coefficients of Determination and Nondetermination – Interpretation and Uses of The Pearson Correlation.

Unit IV: Regression Analysis - I

Regression Model - Goals of Regression Analysis - Statistical Computing in Regression Analysis - Simple Linear Regression – Multiple Linear Regression – Logistic Regression – Poisson Regression

Unit V: Regression Analysis - II

Detection of Outliers and Influential Observations: Detection of Outliers in Multiple Linear Regression - Detection of Influential Observations in Multiple Linear Regression - Test for Mean-shift Outliers - Graphical Display of Regression Diagnosis. Model Selection: Effect of Underfitting and Overfitting - All Possible Regressions – Stepwise Selection. Model Diagnostics: Test Heteroscedasticity - Detection of Regression Functional Form

Text Books:

1. Cheryl Ann Willard, *"Statistical Methods: An Introduction to Basic Statistical Concepts and Analysis"* ,Routledge, 2020. (Unit – I – III)
2. Xin Yan &Xiaogang Su, *"Linear Regression Analysis : Theory and Computing*", World Scientific Pulishing Ltd, 2009. (Unit – IV: Chapter 1,2,3, 8.5,8.6; Unit – V: Chapters 4.2,4.3, 4.4, 4.5, 5.1-5.3, 6.1, 6.2)

Reference Books:

1. John.E.Freund, Irwin Miller, Marylees Miller *"Mathematical Statistics with Applications*", 8th, Prentice Hall of India, 2014
2. Ross, Sheldon. M, *"Introduction to Probability and Statistics for Engineers and Scientists"*, Academic Press, 2009
3. D.C Montgomery, E.A Peck and G.G Vining, *"Introduction to Linear Regression Analysis***"**, John Wiley and Sons, 2003.
4. S. Chatterjee and AHadi, **"***Regression Analysis by Example***"**, 4th Ed., John Wiley and Sons, Inc, 2006