Rajiv Gandhi ProudyogikiVishwavidyalaya Bhopal

M.Tech Computer - Science and Engineering (Data Science)

First Semester Syllabus

MTCD 101-ComputationalLinearAlgebra

CourseObjectives

- Thecoursewilllay downthebasicconceptsand techniquesoflinearalgebraandcalculusneededforsubsequent study.
- Thecoursewillexploretheconceptsinitiallythroughcomputationalexperimentsandt hentrytounderstandthe conceptsand theory behind it.
- Thecoursewillprovideanappreciationofthewideapplicationofthesedisciplineswithi nthescientificworld.

Syllabus

UNIT I

Matrices and Gaussian Elimination – Introduction, geometry of linear equations, Gaussian elimination,matrix multiplication, inverses and transposes.Vector spaces and Linear equations–Vector spaces and sub spaces, linear independence, basis and dimension, four fundamental subspaces.

UNIT II

Orthogonality-Perpendicularvectorsandorthogonal subspaces,innerproductsandprojectionsontolines,projectionsandleastsquareapplications,or thogonalbasis,orthogonalspaces,orthogonalmatrices,GramSchmidtorthogonalization,FFT.

UNIT III

Probability, compound probability and discrete random variable. Binomial, Normal and Poisson's distributions, Sampling distribution, elementary concept of estimation and theory of hypothesis, recurred relations.

UNIT IV

EigenvaluesandEigenvectors– Introduction,diagonalformofamatrix,differenceequationsandthepowersofA^k,PositiveDefin iteMatrices-Minima,maximaandsaddlepoints,testsforpositivedefiniteness,semi-definiteandindefinitematrices,SingularValueDecomposition,Iterativemethods for Ax = b.

UNIT V

Introduction to special matrices - Fourier transforms: discrete and continuous, shiftmatrices and circulant matrices, Kronecker product, sine and cosine transforms from Kronecker sums, Toeplitz matrices and shift in variant filters, graphs and Laplacians and Kirchhoff' slaws, clustering by spectralmethodsandk-means,completingrankonematrices,orthogonalProcrustesproblem,distancematrices

Textbook/ References

Gilbert Strang, Linear Algebra and its Applications, Fourth Edition, Cambridge University Press. 2009. Gene H. Golub and V. Van Loan, Matrix Computations, Third Edition, John Hopkins University Press, Baltimore, 1996.

David C. Lay, Linear Algebra and Its Applications, Pearson Addison Wesley, 2002.

Strang, Gilbert. Linear algebra and learning from data. Cambridge: Wellesley-Cambridge Press, 2019.

MTCD 102-Advanced data structures and Algorithm

Syllabus

UNIT I

INTRODUCTION: Basic concepts of OOPs – Templates – Algorithm Analysis – ADT - List (Singly, Doubly and Circular) Implementation - Array, Pointer, Cursor Implementation

UNIT II

BASIC DATA STRUCTURES: Stacks and Queues – ADT, Implementation and Applications - Trees – General, Binary, Binary Search, Expression Search, AVL, Splay, B-Trees – Implementations - Tree Traversals.

UNIT III

ADVANCED DATA STRUCTURES: Set – Implementation – Basic operations on set – Priority Queue – Implementation - Graphs – Directed Graphs – Shortest Path Problem - Undirected Graph - Spanning Trees – Graph Traversals

UNIT IV

MEMORY MANAGEMENT; Issues - Managing Equal Sized Blocks - Garbage Collection Algorithms for Equal Sized Blocks - Storage Allocation for Objects with Mixed Sizes - Buddy Systems - Storage Compaction

UNIT V

SEARCHING, SORTING AND DESIGN TECHNIQUES: Searching Techniques, Sorting – Internal Sorting – Bubble Sort, Insertion Sort, Quick Sort, Heap Sort, Bin Sort, Radix Sort – External Sorting – Merge Sort, Multi-way Merge Sort, Polyphase Sorting - Design Techniques - Divide and Conquer - Dynamic Programming - Greedy Algorithm – Backtracking - Local Search Algorithms

Reference Books:

1. Mark Allen Weiss, “Data Structures and Algorithm Analysis in C++”, Pearson P
2. Aho, Hopcroft, Ullman, “Data Structures and Algorithms”, Pearson Education P
3. Drozdek, Data Structures and algorithm in Jawa, Cengage (Thomson)
4. Gilberg, Data structures Using C++, Cengage
5. Horowitz, Sahni, Rajasekaran, “Computer Algorithms”, Galgotia,
6. Tanenbaum A.S., Langram Y, Augestien M.J., “Data Structures using C & C++”, Prentice Hall of India, 2002

MTCD 103-Machine Learning

Course Outcomes:

After completing the course student should be able to:

1. Describe in-depth about theories, methods, and algorithms in machine learning.
2. Find and analyze the optimal hyper parameters of the machine learning algorithms.
3. Examine the nature of a problem at hand and determine whether a machine learning can solve it efficiently enough.
4. Solve and implement the real-world problems using machine learning.

Syllabus:

UNIT I

Introduction to machine learning (ML): Basics of ML, History of ML, Evolution of ML, ML Models, Learning and testing models, ML Algorithm and Convergence, MLTechniques, Types of ML, supervised and unsupervised learning, classification and clustering, Applications of ML,Bias-Variance tradeoff.

UNIT II

Neural Networks: McCulloch Pitts Neuron models,Activation Functions, Loss Functions, perceptron,Gradient Descent,Multilayer neural networks: back-propagation, backpropagation calculus, Initialization, Training rules, issues in back-propagation, Bayesian Learning,Competitive learning and self-organization map.

UNIT III

Support Vector Machines(SVM):SVM Formulation, Interpretation & Analysis, hard and soft margin, Hinge loss, SVM dual, SVM tuning parameters, SVM Kernels, twin SVM.

UNIT IV

Clustering: K-Means Clustering, Mean Shift Clustering, Aagglomerative clustering, Association Rule Mining, Partition Clustering, Hierarchical Clustering, Birch Algorithm, CURE Algorithm, Density-based Clustering, Gaussian Mixture Models, and Expectation Maximization. Parameters estimations – MLE,MAP.

UNIT V

Learning Theory: Probably Approximately Correct (PAC) Model, PAC Learnability, Agnostic PAC Learning, Theoretical analysis of machine learning problems and algorithms,Generalization error bounds,VC Model,MLTools.

Recommended Books:

1. Tom Mitchell, Machine Learning, McGraw-Hill, 1997.
2. Leonard Kaufman and P. J. Rousseau. Finding groups in data: An introduction to cluster analysis, Wiley, 2005
3. NelloCristianini and John Shawe-Taylor, An Introduction to Support Vector Machines,Cambridge University Press, 2000.
4. Bernhard Schölkopf and Alexander J. Smola, Learning with Kernels, MIT Press, 2002.
5. Shai Shalev-Shwartz and Shai Ben-David, Understanding Machine Learning:From Theory to Algorithms, Cambridge University Press.,2014

MDSCS104-DataScience

Unit 1: Introduction to core concepts and technologies: Introduction Terminology, data science process, Data science toolkit, Types of data, Example applications.

Unit 2: Data collection and management: Introduction, Sources of data, Data collection and APIs. Exploring and fixing data. Data storage and management, Using multiple data sources.

Unit 3. Data analysis: Introduction , Terminology and concepts. Introduction to statistics Variance ,Distribution properties and arithmetic Samples/CLT, Basic machine learning algorithms ,Linear regression ,SVM, Naive Bayes.

Unit 4: Data Visualization: Introduction ,Types of data visualization, Data for visualization, Data types, Data encodings, Retinal variables, Mapping variables to encodings. Visual encodings.

Unit 5: Applications of Data Science Technologies for visualization, Bokeh (Python) Recent trends in various data collection and analysis techniques various visualization techniques, application development methods of used in data science.

REFERENCE BOOKS

1. Cathy O'Neil and Rachel schutt ,Dong Data Science, Straight Talk from the Frontline. O'Reilly.

2. Jure Leskovek, Anand Rajaraman and Jeffrey Ullman . Mining of Massive Datasets. V2.1. Cambridge University Press

MTCD 105(A) Big Data

UNIT-I Introduction to Big Data. What is Big Data. Why Big Data is Important. Meet Hadoop. Data. Data Storage and Analysis. Comparison with other systems. Grid Computing. A brief history of Hadoop. Apache hadoop and the Hadoop EcoSystem. Linux refresher; VMWare Installation of Hadoop.

UNIT-II The design of HDFS. HDFS concepts. Command line interface to HDFS. Hadoop File systems. Interfaces. Java Interface to Hadoop. Anatomy of a file read. Anatomy of a file write. Replica placement and Coherency Model. Parallel copying with distcp, Keeping an HDFS cluster balanced.

UNIT-III (Introduction. Analyzing data with unix tools. Analyzing data with hadoop. Java MapReduce classes (new API). Data flow, combiner functions, Running a distributed MapReduce Job. Configuration API. Setting up the development environment. Managing configuration. Writing a unit test with MRUnit. Running a job in local job runner. Running on a cluster.Launching a job. The MapReduce WebUI.

UNIT-IV Classic Mapreduce. Job submission. Job Initialization. Task Assignment. Task execution .Progress and status updates. Job Completion. Shuffle and sort on Map and reducer side. Configuration tuning. MapReduce Types. Input formats. Output formats ,Sorting. Map side and Reduce side joins.

UNIT-V The Hive Shell. Hive services. Hive clients. The meta store. Comparison with traditional databases. HiveQl. Hbasics. Concepts. Implementation. Java and Mapreduce clients. Loading data, web queries.

TEXT BOOKS:

1. Tom White, Hadoop, “The Definitive Guide”, 3rd Edition, O’Reilly Publications, 2012.

2. Dirk deRoos, Chris Eaton, George Lapis, Paul Zikopoulos, Tom Deutsch , “Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data”, McGrawHill Osborne Media; 1 edition, 2011.

MTCD 105(B) Data Preparation and Analysis

UNIT- I Data Gathering and Preparation: Data formats, Parsing and transformation, Scalability and real-time issues.

UNIT -II Data Cleaning: Consistency checking, Heterogeneous and missing data, Data Transformation and Segmentation.

UNIT -III Exploratory Analysis: Descriptive and comparative statistics, Clustering and association, Hypothesis Generation.

UNIT- IV Visualization: Designing visualizations, Time series, Geo located data, Correlations and Connections, Hierarchies and networks, interactivity.

UNIT- V Statistics : Descriptive statistics, Central tendency, Variation , Shape, Inferential statistics Confidence intervals, Hypothesis tests, Chi-square, One-way analysis of variance, Comparative statistics

References: 1. Making sense of Data : A practical Guide to Exploratory Data Analysis and Data Mining, by Glenn J. Myatt.,Wiley

MTCD 105(C) Information Retrieval

UNIT-I Introduction - History of IR- Components of IR - Issues -Open source Search engine Frameworks - The Impact of the web on IR - The role of artificial intelligence (AI) in IR – IR Versus Web Search - Components of a search engine, Characterizing the web.

UNIT -II Boolean and Vector space retrieval models- Term weighting - TF-IDF weighting- cosine similarity - Preprocessing - Inverted indices - efficient processing with sparse vectors Language Model based IR - Probabilistic IR -Latent Semantic indexing - Relevance feedback and query expansion.

UNIT- III Web search overview, web structure the user paid placement search engine optimization, Web Search Architectures - crawling - meta-crawlers, Focused Crawling - web indexes – Nearduplicate detection - Index Compression - XML retrieval.

UNIT -IV Link Analysis -hubs and authorities - Page Rank and HITS algorithms -Searching and Ranking - Relevance Scoring and ranking for Web - Similarity - Hadoop & Map Reduce - Evaluation - Personalized search - Collaborative filtering and content-based recommendation of documents And products - handling invisible Web - Snippet generation, Summarization. Question Answering, Cross-Lingual Retrieval.

UNIT -V Information filtering: organization and relevance feedback - Text Mining- Text classification and clustering - Categorization algorithms, naive Bayes, decision trees and nearest neighbor - Clustering algorithms: agglomerative clustering, k-means, expectation maximization (EM).

References:

1. C. Manning, P. Raghvan and H Schutze: Introduction to Information Retrieval, Cambridge University Press, 2008.

2. Ricardo Baeza -Yates and Berthier Ribeiro –Neto, Modern Information Retrieval The Concepts and Technology behind Search 2nd Edition, ACM Press Books 2011.

3. Bruce Croft, Donald Metzler and Trevor Strohman Search Engines Information Retrieval in Practice 1st Edition Addison Wesley, 2009

4.Mark Levene, An Introduction to Search Engines and Web Navigation, 2nd Edition Wiley 2010.

MTCD 105(D) Data Warehousing and Data Mining

UNIT 1 Introduction: Data Mining: Definitions, KDD v/s Data Mining, DBMS v/s Data Mining , DM techniques, Mining problems, Issues and Challenges in DM, DM Application areas. Association Rules & Clustering Techniques: Introduction, Various association algorithms like A Priori, Partition, Pincer search etc., Generalized association rules.

UNIT 2 Clustering paradigms; Partitioning algorithms like K-Medioid, CLARA, CLARANS; Hierarchical clustering, DBSCAN, BIRCH, CURE; categorical clustering algorithms, STIRR, ROCK, CACTUS.

UNIT 3 Other DM techniques & Web Mining: Application of Neural Network, AI, Fuzzy logic and Genetic algorithm, Decision tree in DM. Web Mining, Web content mining, Web structure Mining, Web Usage Mining.

UNIT 4 Temporal and spatial DM: Temporal association rules, Sequence Mining, GSP, SPADE, SPIRIT, and WUM algorithms, Episode Discovery, Event prediction, Time series analysis.

UNIT 5 Spatial Mining, Spatial Mining tasks, Spatial clustering, Spatial Trends, Data Mining of Image and Video: A case study. Image and Video representation techniques, feature extraction, motion analysis, content based image and video retrieval, clustering and association paradigm, knowledge discovery.

Reference Books:

1. Data Mining Techniques; Arun K.Pujari ; University Press.

2. Data Mining; Adriaans&Zantinge; Pearson education.

3. Mastering Data Mining; Berry Linoff; Wiley.

4. Data Mining; Dunham; Pearson education. 5. Text Mining Applications, Konchandy, Cengage