Introduction to Data Science

SYLLABUS

Module – 1 (Introduction to Data Science)

A brief introduction to data – structured, unstructured, semi-structured, data sets & patterns, Brief history of Data Science, Introduction to Data Science, Importance of Data Science, Differences between AI, ML, DL, Data Science & Data Analytics, Real world applications of data science, Steps in data science process

Simple case study based on real life applications such as - Market research case, tracking disease outbreaks, business predictions, (for example, Rating a product design) etc., Ethical and privacy implications of Data Science.

Tools and Skills Needed – brief introduction of platforms, tools, frameworks, languages, databases and libraries, Current trends & major research challenges in data science.

Module – 2 (Data Mining & Preprocessing)

Data Mining, Kinds of data - mining, Data Preprocessing. An Overview - Data Quality, Need to preprocess the data. Major Tasks in Data Preprocessing.

Data cleaning - Missing Values Noisy Data, Data Cleaning as a Process, Data Integration, Data Reduction, Data transformation and Data Discretization. Introduction to Data Visualization

Module - 3 (Classification Models)

Classification - Basic Concepts, Decision Tree Induction, Bayes Classification Methods- Naive Bayesian Classification, Rule-Based Classification

Classification Advanced Methods - Bayesian Belief Networks, Classification by Back propagation, A Multilayer Feed-Forward Neural Network, Back propagation, Support Vector Machines, Lazy Learners, K-Nearest-Neighbour Classifiers, Case-Based Reasoning

Module - 4 (Association Mining and Cluster Analysis)

Mining Frequent Patterns, Associations, and Correlations. Basic Concepts Frequent Itemset Mining Methods, Apriori Algorithm, Generating Association Rules from Frequent Itemsets Cluster Analysis, Partitioning Methods, Hierarchical Methods, Agglomerative versus Divisive Hierarchical Clustering, Distance Measures in Algorithmic Methods. Density-Based Methods - DBSCAN

Module - 5 (Evaluation)

Evaluating model performance-Confusion matrices, Precision and recall, Sensitivity and specificity, F-measure, ROC curves, Cross validation, K-fold cross validation, Bootstrap sampling. Improving model performance - Bagging, Boosting, Random forests.

Text Books

1. Sanjeev J. Wagh, Manisha S. Bhende, and Anuradha D. Thakare, *Fundamentals of Data Science*, CRC press
2. Jiawei Han, Michelin Kamber, Jian Pei, *Data mining Concepts and Techniques*, Third Edition, 2012, Morgan Kaufmann Publishers
3. Brett Lantz, *Machine Learning with R*, Second edition, PackT publishing 2015

Introduction to Machine Learning

SYLLABUS

Module-1 (Overview of machine learning)

Introduction to Machine Learning, Machine learning paradigms-supervised, semi-supervised, unsupervised, reinforcement learning.

Supervised learning- Input representation, Hypothesis class, Version space, Vapnik-Chervonenk is (VC) Dimension, Probably Approximately Correct Learning (PAC), Noise, Learning Multiple classes, Model Selection and Generalization

Module-2 (Supervised Learning)

Dimensionality reduction – Subset selection, Principal Component Analysis.

Regression - Linear regression with one variable, Linear regression with multiple variables, solution using gradient descent algorithm and matrix method, basic idea of overfitting in regression. Linear Methods for Classification- Logistic regression, Naive Bayes, Decision tree algorithm ID3.

Case Study: Develop a classifier for face detection.

Module-3 (Classification Assessment and Neural Networks (NN))

Classification Performance measures - Precision, Recall, Accuracy, F-Measure, Receiver Operating Characteristic Curve(ROC), Area Under Curve AUC. Bootstrapping, Cross Validation.

Perceptron, Neural Network - Multilayer feed forward network, Activation functions (Sigmoid, ReLU, Tanh), Back propagation algorithm.

Module-4 (Parameter estimation & SVM Classifier)

Basics of parameter estimation - Maximum Likelihood Estimation(MLE) and Maximum a Posteriori estimation(MAP). Bias-Variance decomposition.

Support Vector Machines - Introduction, Maximum Margin hyperplanes, Mathematics behind Maximum Margin Classification, soft margin SVM classifier, non-linear SVM, Kernels for learning non-linear functions, polynomial kernel, Radial Basis Function(RBF), Kernel Trick.

Module-5 (Unsupervised Learning)

Ensemble methods, Voting, Bagging, Boosting.

Unsupervised Learning - Clustering Methods -Similarity measures, K-means clustering, Expectation-Maximization for soft clustering, Hierarchical Clustering Methods , Density based clustering.

Operating System
Syllabus

Module I

Introduction: Operating system overview – Operations, Functions, Service – System calls, Types – Operating System structure - Simple structure, Layered approach, Microkernel, Modules – System boot process.

Module II

Processes - Process states, Process control block, threads, scheduling, Operations on processes - process creation and termination – Inter-process communication - shared memory systems, Message passing systems.

Process Scheduling – Basic concepts- Scheduling criteria -scheduling algorithms- First come First Served, Shortest Job Firs, Priority scheduling, Round robin scheduling

Module III

Process synchronization- Race conditions – Critical section problem – Peterson's solution, Synchronization hardware, Mutex Locks, Semaphores, Monitors – Synchronization problems - Producer Consumer, Dining Philosophers and Readers-Writers.

Deadlocks: Necessary conditions, Resource allocation graphs, Deadlock prevention, Deadlock avoidance – Banker's algorithms, Deadlock detection, Recovery from deadlock.

Module IV

Memory Management: Concept of address spaces, Swapping, Contiguous memory allocation, fixed and variable partitions, Segmentation, Paging. Virtual memory, Demand paging, Page replacement algorithms.

Module V

File System: File concept - Attributes, Operations, types, structure – Access methods, Protection. File-system implementation, Directory implementation. Allocation methods.

Storage Management: Magnetic disks, Solid-state disks, Disk Structure, Disk scheduling, Disk formatting.

Text Book

Abraham Silberschatz, Peter Baer Galvin, Greg Gagne, ' Operating System Concepts' 9th Edition, Wiley India 2015.

Syllabus **Database Management Systems**

Module 1: Introduction & Entity Relationship (ER) Model

Concept & Overview of Database Management Systems (DBMS) - Characteristics of Database system, Database Users, structured, semi-structured and unstructured data. Data Models and Schema - Three Schema architecture. Database Languages, Database architectures and classification.

ER model - Basic concepts, entity set & attributes, notations, Relationships and constraints, cardinality, participation, notations, weak entities, relationships of degree 3.

Module 2: Relational Model

Structure of Relational Databases - Integrity Constraints, Synthesizing ER diagram to relational schema

Introduction to Relational Algebra - select, project, cartesian product operations, join - Equi-join, natural join. query examples, introduction to Structured Query Language (SQL), Data Definition Language (DDL), Table definitions and operations -- CREATE, DROP, ALTER, INSERT, DELETE, UPDATE.

Module 3: SQL DML (Data Manipulation Language), Physical Data Organization

SQL DML (Data Manipulation Language) - SQL queries on single and multiple tables, Nested queries (correlated and non-correlated), Aggregation and grouping, Views, assertions, Triggers, SQL data types.

Physical Data Organization - Review of terms: physical and logical records, blocking factor, pinned and unpinned organization. Heap files, Indexing, Singe level indices, numerical examples, Multi-level-indices, numerical examples, B-Trees & B+-Trees (structure only, algorithms not required), Extendible Hashing, Indexing on multiple keys – grid files.

Module 4: Normalization

Different anomalies in designing a database, The idea of normalization, Functional dependency, Armstrong's Axioms (proofs not required), Closures and their computation, Equivalence of Functional Dependencies (FD), Minimal Cover (proofs not required). First Normal Form (1NF), Second Normal Form (2NF), Third Normal Form (3NF), Boyce Codd Normal Form (BCNF), Lossless join and dependency preserving decomposition, Algorithms for checking Lossless Join (LJ) and Dependency Preserving (DP) properties.

Module 5: Transactions, Concurrency and Recovery, Recent Topics

Transaction Processing Concepts - overview of concurrency control, Transaction Model, Significance of concurrency Control & Recovery, Transaction States, System Log, Desirable Properties of transactions.

Serial schedules, Concurrent and Serializable Schedules, Conflict equivalence and conflict serializability, Recoverable and cascade-less schedules, Locking, Two-phase locking and its variations. Log-based recovery, Deferred database modification, check-pointing.

Introduction to NoSQL Databases, Main characteristics of Key-value DB (examples from: Redis), Document DB (examples from: MongoDB)

Main characteristics of Column - Family DB (examples from: Cassandra) and Graph DB (examples from : ArangoDB)

Data Structures

SYLLABUS

Module 1

Basic Concepts of Data Structures

System Life Cycle, Algorithms, Performance Analysis, Space Complexity, Time Com
Asymptotic Notation, Complexity Calculation of Simple Algorithms

Module 2

Arrays and Searching

Polynomial representation using Arrays, Sparse matrix, Stacks, Queues-Circular Queues,
Queues, Double Ended Ques, Evaluation of Expressions
Linear Search and Binary Search

Module 3

Linked List and Memory Management

Self Referential Structes, Dynamic Memory Allocation, Singly Lin   ked List-Operat
Linked List. Doubly liked List, Circular Linked List, Stacks and Qu   eues using Link
Polynomial representation using Linked List
Memory allocation and deallocation-First-fit, Best-fit and Worst-fit allo   cation schemes

Module 4

Trees and Graphs

Trees, Binary Trees-Tree operations, Binary Tree Representation, Tree T   raversals, Binar
Trees- Binary Search Tree Operations
Graphs, Representation of Graphs, Depth First Search and Breadth   First Search on
Applications of Graphs

Module 5

Sorting and Hashing

Sorting Techniques – Selection Sort, Insertion Sort, Quick Sort, Merge Sort and Heap So
Hashing- Hashing Techniques, Collision Resolution, Overflow handling, Hashing fun
Mid square, Division, Folding, Digit Analysis

Text Book

1. Ellis Horowitz, Sartaj Sahni and Susan Anderson-Freed, Universities
   Fundamentals of Data Structures in C