JSS MAHAVIDYAPEETHA
JSS SCIENCE & TECHNOLOGY UNIVERSITY

Sri Jayachamarajendra College of Engineering
Mysuru-570006.

Department of Information Science & Engineering

Master of Technology
In
Data Science

SCHEME

I to IV semesters

2017-2018

Scheme of Teaching and Examination

M.Tech. in Data Science

First Semester M.Tech. (DS) 2017-2018

SL. No.	Subject Code	Course Title	Teaching Department	Credits				Contact Hours	Marks			Exam Duration (Hrs)
				L	T	P	Total		CIE	SEE	Total	
1.	SDS110	Principles of Data Science	IS&E	4	1	0	5.0	6	50	50	100	3
2.	SDS120	Big Data Analytics	IS&E	4	1	0	5.0	6	50	50	100	3
3.	SDS130	Machine Learning	IS&E	4	0	1	5.0	6	50	50	100	3
4a.	SDS141	Computational Statistical Methods	IS&E	4	0	1	5.0	6	50	50	100	3
4b.	SDS142	Information Retrieval	IS&E	4	0	1	5.0	6	50	50	100	3
4c.	SDS143	Image & Video Analytics	IS&E	4	0	1	5.0	6	50	50	100	3
5a.	SDS151	Numerical Linear Algebra	IS&E	4	0	1	5.0	6	50	50	100	3
5b.	SDS152	System Security	IS&E	4	0	1	5.0	6	50	50	100	3
5c.	SDS153	Exploratory Data Analysis & Visualization	IS&E	4	0	1	5.0	6	50	50	100	3
6.	SDS160	Minor Project – I	IS&E	0	0	1.5	1.5	3	50	-	50	-
7.	SDS170	Seminar – I	IS&E	0	1.5	0	1.5	3	50	-	50	-
Total				20	3.5	4.5	28	36	350	250	600	-

Scheme of Teaching and Examination

M.Tech. in Data Science

Second Semester M.Tech. (DS) 2017-2018

SL. No.	Subject Code	Course Title	Teaching Department	Credits				Contact Hours	Marks			Exam Duration (Hrs)
				L	T	P	Total		CIE	SEE	Total	
1.	SDS210	Advanced Data Mining Techniques	IS&E	4	1	0	5.0	6	50	50	100	3
2.	SDS220	Scalable Systems for Data Science	IS&E	4	0	1	5.0	6	50	50	100	3
3.	SDS230	Deep Learning	IS&E	4	0	1	5.0	6	50	50	100	3
4a.	SDS241	Optimization Theory	IS&E	4	1	0	5.0	6	50	50	100	3
4b.	SDS242	Computational Linguistics	IS&E	4	1	0	5.0	6	50	50	100	3
4c.	SDS243	Bioinformatics	IS&E	4	1	0	5.0	6	50	50	100	3
5a.	SDS251	Cloud Computing & Virtualization	IS&E	4	0	1	5.0	6	50	50	100	3
5b.	SDS252	Web Databases & Information Systems	IS&E	4	0	1	5.0	6	50	50	100	3
5c.	SDS253	Social & Information Network Analysis	IS&E	4	0	1	5.0	6	50	50	100	3
6.	SDS260	Minor Project – II	IS&E	0	0	1.5	1.5	3	50	-	50	-
7.	SDS270	Seminar – II	IS&E	0	1.5	0	1.5	3	50	-	50	-
Total				20	3.5	4.5	28	36	350	250	600	-

Scheme of Teaching and Examination

MTech in Data Science

Third Semester MTech (DS) 2017-2018

Sl.No.	**Subject Code**	**Course title**	**Teaching Department**	**Credits**				**Contact Hours**	**Marks**			**Exam Duration**
				L	**T**	**P**	**Total**		**CIE**	**SEE**	**Total**	
1	SDS31T	Practical Training in Industry/Exploration in Research	IS&E	--	--	4	4	-	100	-	100	-
2	SDS32P	Project Work (Phase – I)	IS&E	--	--	10	14	-	100	-	100	-
				Total Credits			**18**		**Total Marks**		**200**	

Scheme of Teaching and Examination

M.Tech. in Data Science

Fourth Semester M.Tech. (DS) 2017-2018

Sl.No.	Subject Code	Course title	Teaching Department	Credits				Contact Hours	Marks			Exam Duration
				L	T	P	Total		CIE	SEE	Total	
1	SDS41P	Project Work (Phase – II)	IS&E	--	--	26	26	-	100	200	300	-
				Total Credits			**26**	**-**	**Total Marks**		**300**	-

SDS110 PRINCIPLES OF DATA SCIENCE

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction 12 Hours

What is Data Science?, Basic Terminology, Why Data Science?, Example – Sigma Technologies, The data science Venn diagram, The math: Example – Spawner-Recruit Models, Computer programming (Preferably Phyton/R/Matlab/PERL), Some more terminology, Some Data science case studies, Data Models and its types.

Types of Data, Flavors of data, Structured versus unstructured data, Quantitative versus qualitative data, The four levels of data, The Five Steps of Data Science.

Basic Mathematics – Vectors and Matrices 14 Hours

Vectors and Linear Combinations, Lengths and Dot Products, Matrices, Solving Linear Equations: Vectors and Linear Equations, The Idea of Elimination, Elimination Using Matrices, Rules for Matrix Operations, Inverse Matrices, Elimination = Factorization: A = LU, Transposes and Permutations, Vector Spaces and Subspaces: Spaces of Vectors, The Nullspace of A: Solving Ax = 0 and Rx = 0, The Complete Solution to Ax = b, Independence, Basis and Dimension, Dimensions of the Four Subspaces.

Probability and Statistics 14 Hours

Basic definitions, Probability, Bayesian versus Frequentist, Frequentist approach, Compound events, Conditional probability, The rules of probability, Collectively exhaustive events, Bayesian ideas revisited, Bayes theorem, Random variables

Basic Statistics, What are statistics?, How do we obtain and sample data?, Obtaining data, Sampling data, How do we measure statistics?, Point estimates, Sampling distributions, Confidence intervals, Hypothesis tests, Conducting a hypothesis test, Type I and type II errors, Hypothesis test for categorical variables

Visualization 10 Hours

Basic principles, ideas and tools for data visualization, why does communication matter?, Identifying effective and ineffective visualizations, Scatter plots, Line graphs, Bar charts, Histograms, Box plots, When graphs and statistics lie, Correlation versus causation, Simpson's paradox, If correlation doesn't imply causation, then what does?, Verbal communication, The why/how/what strategy of presenting, Data Science ethical issues.

Text Books:

1. Principles of Data Science, Sinan Ozdemir, PACKT Publisher, First Edition, 2016.
2. Introduction to Linear Algebra, Gilbert Strang, Wellesley-Cambridge Press, Fifth Edition, 2016.

Reference Books:

1. Doing Data Science: Straight Talk from the Frontline, Cathy O'Neil, Rachel Schutt, O'Reilly Media, 2013.
2. Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman, Second Edition, Cambridge University Press Publisher, 2015.
3. Materials from Internet Sources

SDS120 BIG DATA ANALYTICS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction & Perspective of Big Data **10 Hours**

Overview of Big Data, History, Structuring Big Data, Types of Data, Elements of Big Data, Data analytics project life cycle, Problems & challenges in understanding Data Analytics, Web page categorization (In detail), Case studies on: Stock Market changes.
Application of Data Analytics in Digital market, Big Data benefit areas, Various Analytical approaches, Cross Channel Life cycle marketing, Use of Big Data in Social Networking, Use of Big Data in Business Intelligence, Use of Big Data in preventing Fraudulent activities, Use of Big Data in Retail Industry, Use of RFID Data in Retail, Big Data in Health Care, Predictive and Disruptive Analytics, Content delivery and market optimization.

Big Data Technology **10 Hours**

Exploring Big Data Stack, Virtualization, Virtualization Approaches, Distributed and parallel computing for Big Data, Introducing Hadoop, Hadoop Ecosystem, Hadoop Distributed File Systems(HDFS), Features of HDFS : Hadoop YARN, MAP Reduce, Features of Map Reduce, Working of Map Reduce, Techniques to Optimize Map Reduce Jobs, Uses of Map Reduce, HBase, Features of HBase, Role of HBase in Big Data processing, Other tools of Hadoop (Hive, Pig and Pig Latin, Sqoop, ZooKeeper, Flume, OOZie), The cloud and Big Data, Cloud Deployment Models, Cloud Delivery Models, Cloud providers in Big Data Market.

Mining Data Streams **12 Hours**

The Stream Data Model, A Data-Stream-Management System, Examples of Stream Sources, Stream Queries, Issues in Stream Processing, Sampling Data in a Stream, Filtering Streams, Estimating Moments, Dealing With Infinite Streams, Counting Ones in a Window, The Market Basket Analysis, A Priori Algorithm, Handling Larger Datasets in Main Memory, Limited-Pass Algorithms, Counting Frequent Items in a Stream.

Advanced Analytical Theory and Methods **08 Hours**

Analytics on Text, Image, Video, Web, Social Network (A Case Studies on all the different types of Data), Time Series Analysis, NoSQL, Recommendation System: A Model, Content Based Recommendations, Collaborative Filtering, Dimensionality Reduction Problem, The NetFlix Problem.

Large Scale Machine Learning **10 Hours**

Introduction, Types of Machine Learning Algorithms, Machine Learning Architecture, Applications of Machine Learning, Supervised Machine Learning Algorithms (Problems on Classification): Bayseian Networks, Learning from Nearest Neighbors, Decision Trees, Support Vector Machines, Neural Networks, Unsupervised Machine Learning Algorithms (Problems on Clustering): Hierarchical Clustering Techniques, Partitional Clustering techniques, Distance measures.

Text Books:

1. Big Data: Black Book, DT Editorial Services, Dream Tech Press Publishers, 2015.
2. Mining of Massive Datasets, Jure Leskovec, Anand Rajaraman, Jeff Ullman, Second Edition,
   Cambridge University Press Publisher, 2015.

Reference Books:

1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015.
2. Selected Research Articles from Internet.

SDS130 MACHINE LEARNING

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction & Bayesian Decision Theory **10 Hours**

What Is Machine Learning?, Challenges, Examples of Machine Learning Applications, Present Research Avenues, Introduction to Bayesian Decision Theory, Classification, Losses and Risks, Discriminant Functions, Utility Theory, Association Rules

Dimensionality Reduction **10 Hours**

Introduction, Feature Generation, Feature Selection, Principal Component Analysis, Factor Analysis, Multidimensional Scaling, Linear Discriminant Analysis, Locality Preserving Projections (LPP) and it's variants, Locality Preserving Indexing and its variants.

Supervised Learning **12 Hours**

Learning a Class from Examples, Probably Approximately Correct (PAC) Learning, Noise, Learning Multiple Classes, Regression, Model Selection and Generalization, Dimensions of a Supervised Machine Learning Algorithms, Decision Tree Induction, Nearest Neighbors, Bayesian Classifier, Artificial Neural Networks, Model Over fitting, Performance Evaluation of classifiers.

Clustering **10 Hours**

Basic Concepts, Proximity Measures, Sequential Algorithms, Hierarchical Algorithms, Schemes based on Functional Optimization, Clustering Algorithms based on Graph Theory, Cluster Validity.

Machine Learning Applications in Software Engineering **8 Hours**

The challenges, Related Issues, Learning Approaches, SE tasks for ML Applications, State of the Practice in ML & SE, Present Status, Applying ML algorithms to SE Tasks.

Text Books:

1. Introduction to Machine Learning, Ethem Alpaydin, Second Edition, PHI Learning Publisher, 2013 edition.
2. Pattern Recognition, Sergios Theodoridis and Konstantinos Koutroumbas, Fourth Edition, Academic Press Publisher, 2014.

Reference Books:

1. Machine Learning, Tom M. Mitchell, Mc Graw Hil Publishers, 1997.
2. Machine Learning Applications in Software Engineering, Du Zhang and Jeffrey J. P. Tsai, World Scientific Publishers, 2005.
3. Pattern Recognition and Machine Learning, Christopher M. Bishop, Spriger Publishers, 2011.
4. Related Research Articles

SDS141 COMPUTATIONAL STATISTICAL METHODS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Statistical Learning **10 hours**

Introduction, What is Statistical Learning, Assessing Model Accuracy

Linear Regression **10 hours**

Simple Linear Regression, Multiple Linear Regression, Other considerations in the Regression Model, The Market Plan, Comparison of Linear Regression with K-Nearest Neighbors

Classification **10 hours**

Overview of Classification, Why not Linear Regression, Logistics Regression, LDA, Comparison of Classification Methods

Resampling Methods/ Linear Model Selection & Regularization **10 hours**

Cross Validation, The Bootstrap

Subset Selection, Shrinkage Methods, Dimensionality Reduction Methods, Considerations in High Dimension

Tree based Methods/SVM/Unsupervised Learning **10 hours**

Basics of Decision Trees, Bagging, Random Forests, Boosting

SVMs: Maximal Margin Classifier, Support Vector Classifiers, SVMs, SVMs with more than two classes, Relation to Logistic Regression

Unsupervised Learning: Challenges, PCA, Cluster Methods

Text Books:

1. Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, “An Introduction to Statistical Learning with Applications in R”, 2014, Springer.
2. Geof Givens, Jennifer Hoeting, “Computational Statistics”, 2nd edition, 2013, Wiley
3. Max Kuhn, Kjell Johnson, “Applied Predictive Modeling”, 2013, Springer

SDS142 INFORMATION RETRIEVAL

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 hours**

Motivation, Basic concepts, Past, present, and future, The retrieval process.
Modeling: Introduction, A taxonomy of information retrieval models, Retrieval: Adhoc and filtering, A formal characterization of IR models, Classic information retrieval, Alternative set theoretic models, Alternative algebraic models, Alternative probabilistic models, Structured text retrieval models, Models for browsing.

Evaluation **10 hours**

Retrieval Evaluation: Introduction, Retrieval performance evaluation, Reference collections.
Query Languages: Introduction, keyword-based querying, Pattern matching, Structural queries, Query protocols.
Query Operations: Introduction, User relevance feedback, Automatic local analysis, Automatic global analysis.

Properties **10 hours**

Text and Multimedia Languages and Properties: Introduction, Metadata, Text, Markup languages, Multimedia.
Text Operations: Introduction, Document preprocessing, Document clustering, Text compression, comparing text compression techniques.

Indexing & Searching **10 hours**

Introduction; Inverted Files; Other indices for text; Boolean queries; Sequential searching; Pattern matching; Structural queries; Compression.
Parallel and Distributed IR: Introduction, Parallel IR, Distributed IR.

Interface & Visualization **10 hours**

User Interfaces and Visualization: Introduction, Human-Computer interaction, The information access process, Starting pints, Query specification, Context, Using relevance judgments, Interface support for the search process.
Searching the Web: Introduction, Challenges, Characterizing the web, Search engines, Browsing, Metasearchers, Finding the needle in the haystack, Searching using hyperlinks.

Text Books:

1. Ricardo Baeza-Yates, Berthier Ribeiro-Neto: Modern Information Retrieval, Pearson, 1999.

Reference Books:

1. David A. Grossman, Ophir Frieder: Information Retrieval Algorithms and Heuristics, 2nd Edition, Springer, 2004.

SDS143 IMAGE & VIDEO ANALYTICS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 Hours**

Digital Image Fundamentals, Elements of visual perception, Image sensing and acquisition, Sampling and Quantization, Relationships between pixels, Linear and Non Linear operations.

Image and Video Processing: Basic Linear Filtering to Image Enhancement, Non Linear Filtering for Image Analysis and Enhancement, Morphological Filtering for Image Enhancement and Detection, Image Restoration, Motion Detection and Estimation, Video enhancement and restoration.

Image Enhancement **12 Hours**

Basic Gray Level Transformations, Histogram Processing, Enhancement using Arithmetic and Logical Operations, Basics of Spatial Filtering, Image Enhancement in Frequency Domain, Smoothing and Sharpening Frequency Domain Filters, Homomorphic Filters.

Digital Video & Motion Estimation **8 Hours**

Human Visual System and Color, Analog and Digital Video, 3D Video, Digital Video Applications, Image and Video quality, Motion Models, 2D Apparent Motion estimation, Differential Methods, Matching Methods, Non Linear Optimization Methods, Transform Domain Methods.

UNIT IV

Image and Video Segmentation **10 Hours**

Detection of Discontinues, Edge Linking and Boundary Detection, Threshold Based Segmentation, Region Based Segmentation, Segmentation by Morphological Watersheds, The use of Motions in Segmentation, Change Detection, Motion Tracking, Image and Video Matting.

Image & Video Compression **10 Hours**

Basics of Image Compression, Lossless Image Compression, Discrete Cosine Transform Coding and JPEG, Wavelet Transform Coding and JPEG2000, Video Compression Approaches, Early Video Compression Standards, MPEG-4, H.264 Standard, High Efficiency Video Coding (HEVC) Standard, Scalable Video Compression.

Text Books:

1. Digital Video Processing, *Murat Tekalp*, Second Edition, Prentice Hall, 2015.
2. Digital Image Processing, *Rafael C. Gonzalez and Richard E. Woods,* Fourth Edition, Pearson Publisher, 2017.

Reference Books:

1. Handbook of Image and Video Processing, *Alan C. Bovik*, Second Edition, Academic Press, 2005.
2. Various Research articles to discuss Applications.

SDS151 NUMERICAL LINEAR ALGEBRA

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 Hours**

Fundamentals: Matrix –Vector Multiplication, Orthogonal Vectors & Matrices, Norms, The Singular Value Decomposition, More on the SVD.
QR Factorization & Least Squares: Projectors, QR Factorization, Gram – Schmidt Orthogonalization, MATLAB, Householder Triangularization, Least Squares Problems.

Conditioning & Stability **10 Hours**

Conditioning & Condition Numbers, Floating Point Arithmetic, Stability, More on Stability, Stability of Householder Triangularization, Stability of Back Substitution, Conditioning of Least Squares Problems, Stability of Least Squares Algorithms.

Systems of Equations **10 Hours**

Gaussian Elimination, Pivoting, Stability of Gaussian Elimination, Cholesky Factorization.

Eigenvalues **10 Hours**

Eigenvalue Problems, Overview of Eignvalue Algorithms, Reduction to Hessenberg or Tridiagonal form, Rayleigh Quotient, Inverse Iteration, QR Algorithm without Shifts, QR Algorithm with Shifts, Other Eigenvalue Algorithms, Computing SVD.

Iterative Methods **10 Hours**

Overview of Iterative Methods, The Arnoldi Iteration, How Arnoldi Locates Eigenvalues, GMRES, The Lanczos Iteration, From Lanczosto Gauss Quadrature, Conjugate Gradients, Biorthogonalization Methods, Preconditioning.

Text Book:

1. Numerical Linear Algebra, Llyod N Trefethen & Davis Bau III, SIAM.

Reference Book:

1. Numerical Linear Algebra, William Layton and Myron Sussman, University of Pittsburgh Pittsburgh, Pennsylvania, ISBN 978-1-312-32985-0

SDS152 SYSTEM SECURITY

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Security / Cryptography **10 hours**

Is There a Security Problem in Computing? What Does Secure Mean? Attacks, The Meaning of Computer Security, Computer Criminals, Methods of Defense.
Elementary Cryptography, Terminology and Background, Substitution Ciphers, Transpositions (Permutations), Making a Good Encryption Algorithms, The Data Encryption Standard, The AES Encryption Algorithm, Public Key Encryption, The Uses of Encryption, Summary of Encryption

Program / OS Security **10 hours**

Program Security, Secure Programs, Nonmalicious Program Errors, Viruses and Other Malicious Code, Targeted Malicious Code, Controls Against Program Threats, Summary of Program Threats and Controls
Protection in General-Purpose Operating Systems, Protected Objects and Methods of Protection, Memory and Address Protection, Control of Access to General Objects, File Protection Mechanisms, User Authentication, Summary of Security for Users

Trusted OS / Database Security **10 hours**

Designing Trusted Operating Systems, What Is a Trusted System? Security Policies, Models of Security, Trusted Operating System Design, Assurance in Trusted Operating Systems, Summary of Security in Operating Systems
Database and Data Mining Security, Introduction to Database, Security Requirements, Reliability and Integrity, Sensitive Data, Inference, Multilevel Databases, Proposals for Multilevel Security, Data Mining

Network / Administrative Security **10 hours**

Security in Networks, Network Concepts, Threats in Networks, Network Security Controls Section, Firewalls, Intrusion Detection Systems, Secure E-Mail, Summary of Network Security
Administering Security, Security Planning, Risk Analysis, Organizational Security Policies, Physical Security

Cyber Security / Privacy **10 hours**

The Economics of Cyber security, Making a Business Case, Quantifying Security, Modeling Cyber security, Current Research and Future Directions, Summary

Privacy in Computing, Privacy Concepts, Privacy Principles and Policies, Authentication and Privacy, Data Mining, Privacy on the Web,. E-Mail Security, Impacts on Emerging Technologies

Textbook:

1. “Security in Computing”, 4th edition, Charles P. Pfleeger - Pfleeger Consulting Group, Shari Lawrence Pfleeger - RAND Corporation

SDS153 EXPLORATORY DATA ANALYSIS & VISUALIZATION

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **08 hours**

EDA Introduction, What is EDA? EDA vs Classical & Bayesian, EDA vs Summary, EDA Goals, The Role of Graphics, An EDA/Graphics Example, General Problem Categories.

EDA Assumptions **08 hours**

Underlying Assumptions, Importance, Techniques for Testing Assumptions, Interpretation of 4-Plot, Consequences.

EDA Techniques **08 hours**

Introduction, Analysis Questions, Graphical Techniques: Alphabetical, Graphical Techniques: By Problem Category, Quantitative Techniques, Probability Distributions.

EDA Case Studies **12 hours**

Case Studies Introduction, Case Studies : Normal random numbers, Uniform random numbers, Random walk, Josephson Junction Cryothermometry, Beam Deflections, .Filter Transmittance, Standard Resistor, Heat Flow Meter 1, Airplane Glass Failure Time, Ceramic Strength.

Data Visualization **12 hours**

Introduction to R, Rstudio, and Data cleaning and aggregation, Design principles for charts and graphs, ggplot2 and Tableau tools for creating data visualizations, The process creating visualizations and selecting the appropriate visual display, Designing effective digital presentations, Visualization as exploration, Visualizing categorical data, Visualizing time series data, Visualizing multiple variables, Visualizing geospatial data, Dashboard design, Web-based visualizations, Interactive visualizations and motion.

Reference Books:

1. Engineering Statistics Handbook
2. Exploratory Data Analysis With R, Roger D.Peng
3. Interactive Data Visualization for the Web, Scott Murray
4. Advanced Analytics with R and Tableau, Jen Stirrup, Packt Publications
5. https://poldham.github.io/ggplot_pizza_patents_part2j/
6. https://www.tableau.com/solutions/topic/r

SDS210 ADVANCED DATA MINING TECHNIQUES

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 Hours**

The Data Mining Process: Basic Data Types, The Major Building Blocks: A Bird’s Eye View, Scalability Issues and the Streaming Scenario, A Stroll through some Application Scenarios, Data Preparation, Feature Extraction and Portability, **Data Cleaning:** Data Reduction and Transformation, **Similarity and Distances:** Multidimensional Data, Text Similarity Measures, Temporal Similarity Measures, Graph Similarity Measures, Supervised Similarity Functions

Mining Data Stream **10 Hours**

Mining Time-Series Data, Mining Sequence Patterns in Transactional Databases, Mining Sequence Patterns in Biological Data, Graph Mining, Social Network Analysis, Multirelational Data Mining, Multidimensional Analysis and Descriptive Mining of Complex Data Objects, Spatial Data Mining, Multimedia Data Mining, Text Mining, Mining the World Wide Web.

Advanced Concepts in Association Analysis **8 Hours**

Frequent Itemset Generation, Compact Representation of Frequent Itemsets, FP- Growth Algorithms, Handling Categorical and Continuous Attributes, Handling a Concept Hierarchy, Sequntial Patterns, Subgraph Patterns, Infrequent Patterns, Counting Frequent Items in a Stream .

Data Mining Methods as Tools **12 Hours**

Memory-Based Reasoning Methods, Fuzzy Sets in Data Mining, Rough Sets, Support Vector Machines, Genetic Algorithm Support to Data Mining, Performance Evaluation for Predictive Modeling.

Applications and Research Trends in Data Mining **10 Hours**

Data Mining Applications (Financial Data Analysis, Retail Industry, Telecommunication Industry, Biological Data Analysis, Other Scientific Applications, Intrusion Detection), Data Mining System Products and Research Prototypes, Statistical Data Mining, Visual and Audio Data Mining, Data Mining and Collaborative Filtering, Data Mining, Privacy, and Data Security, Trends in Data Mining, Present Research Avenues.

Text Books:

1. Data Mining: Concepts and Techniques, *Jiawei Han, Micheline Kamber, Jian Pei Professor*, Third Edition, Morgan Kauffmann Publishers, 2011.
2. Advanced Data Mining Techniques, *David L. Olson, Dursun Delen*, Springer Publisher, 2008

Reference Books:

1. Data Mining: The Textbook, *Charu C. Aggarwal*, First Edition ,Springer Publisher, 2016.
2. Data Mining: Introductory and Advanced Topics, *Dunham*, First Edition, Pearson Education India Publisher, 2006

SDS220 SCALABLE SYSTEMS FOR DATA SCIENCE

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Big Data **10 hours**

Big Data & Platform Design Goals, Big Data & other computing platforms. Programming for Large Datasets:. Distributed systems, scalability and metrics, Degrees of Parallelism, MapReduce- Uses, Model working, simple and advanced applications programming.Runtime Systems: Hadoop- Open cloud server, Class cluster, Hadoop distributed file system. Hadoop YARN, Hadoop Mapreduce, Fault Tolerance.

Prediction **10 hours**

Prediction over graphs – Scalable learning and inference over graphs, Semi supervised learning (SSL)-self training and co training, Graph Based SSL. Streaming Naive Bayes – Introduction to Naïve Bayes, Complexity of Naïve Bayes, Implementation of Naïve Bayes Classifier, Large vocabulary counting, Sorting – Merge sort, Unix sort, Large vocabulary Naïve Bayes, Distributed Counting, optimizations.

Learning **10 hours**

Scalable Logistic Regression and SGD. Learning as optimization, Stochastic gradient descent, SGD versus streaming, Logistic regression versus Rocchio and Naïve Bayes, Efficient Logistic Regression with Stochastic Gradient Descent, Regularized logistic regression, Sparse updates for Regularized logistic regression, Bounded memory logistic regression, SGD implementation.

Matrix Factorization **10 hours**

Large-scale Matrix Factorization (MF) - Recovering Latent factor in a matrix, Matrix factorization for collaborative filtering, MF for image and text modeling, Large scale MF for distributed SGD, Distributed SGD for Mapreduce.MR Advanced Topics: Inverted Index, PageRank, Distributed graph procesing: Apache Giraph, GoFFish.

Distributed stream processing **10 hours**

Distributed and fault-tolerant realtime computation, Distributed Stream Processing systems- Apache Storm. Parameter Server- Architecture, Key- value vectors, Range Push and Pull, User defined functions on the server, Asynchronous tasks and dependency, Flexible consistency, user defined filters, messages, consistent hashing, server management, worker management. Evaluation of parameter server - Sparse Logistic Regression and Latent Dirichlet Allocation.

Text Books:

1. Select chapters from Mining of Massive Datasets, JureLeskovec, AnandRajaraman and Jeff Ullman, 2nd Edition (v2.1), 2014.
2. Select chapters from Data-Intensive Text Processing with MapReduce, Jimmy Lin and Chris Dyer, 1st Edition, Morgan & Claypool Publishers, 2010
3. Research papers and articles - [MR for ML on Multicore, NIPS 06], [Hogwild!], [Bottou, 2010], [Gemulla et al., KDD 2011]

Reference Books:

1. Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data, EMC Education Services, 2015.
2. Selected Research Articles from Internet.

SDS230 DEEP LEARNING

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Deep Networks Regularization & optimization **10 hours**

Feed forward networks- Gradient based learning, hidden units, backpropagation. Regularization – parameter norm, Dataset augmentation, Noise robustness, semi-supervised learning, multitask learning, early stopping, sparse representation, bagging, ensemble, dropout, manifold learning. Optimization for training deep models- challenges in neural network optimization, adaptive learning rates, and optimization strategies.

Convolution networks **10 hours**

Convolution network, pooling, structured output, data types, efficient convolution algorithm, randomized and unsupervised features, Recurrent and recursive networks- unfold computation graphs, recurrent neural networks, encoder-decoder, deep recurrent network, recursive neural network, echo state network, optimization, and challenges. Practical methodology and its application- performance metrics, selecting hyper parameters. Some application of deep learning like computer vision, speech recognition.

Linear factor models **10 hours**

Probabilistic PCA and factor analysis, independent component analysis, slow feature analysis, sparse coding, and manifold interpretation of PCA. Auto encoders- auto encoders, regularized auto encoders, stochastic auto encoder- decoder, learning manifold with auto encoder, predictive sparse decomposition.

Representation learning **10 hours**

Greedy unsupervised pre-training, transfer learning, distribution representation, exponential gain, providing clues for underlying causes. Structured probabilistic model for deep learning – challenges of unstructured modeling, using graph to describe unstructured model, sampling from graphical models, learning about dependencies, deep learning approach towards structured probabilistic model. Monte carlo methods- sampling monte- carlo methods, importance sampling, markov chain montecarlo methods, gibbs sampling.

Deep generative models **10 hours**

Boltzmann machine, restricted Boltzmann machine, deep belief networks, Boltzmann machine for real valued data, convolutional Boltzmann machine, other Boltzmann machine, back propogation through random operations, directed generative methods, generative stochastic methods, evaluating generative methods.

Text Books:

1. Deep learning - Ian Goodfellow and YoshuaBengio and Aaron Courville, MIT press, Cambridge, Massachusetts,London, ,2016

Reference Books:

1. Fundamentals of Deep Learning:Nikhil Buduma, Nicholas Locascio,O’Reilly media ,2017
2. Deep Learning: Methods and Applications, Li Deng & Dong Yu, 2014.
3. Grokking Deep Learning– Andrew W trask, 2016

SDS241 OPTIMIZATION THEORY

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction 10 Hours

Introduction to Optimization: EIntroduction, Historical Development, Engineering Applications of Optimization' Statement of an Optimization Problem, Classification of Optimization Problems.

Classical Optimization Techniques: Single-Variable Optimization, Multivariable Optimization with No Constraints, Multivariable Optimization with Equality Constraints, Multivariable Optimization with Inequality Constraints, Convex Programming Problem.

Linear Programming 10 Hours

Applications of Linear Programming, Standard Form of a Linear Programming Problem, Geometry of Linear Programming Problems, Definitions and Theorems, Solution of a System of Linear Simultaneous Equations, Pivotal Reduction of a General System of Equations, Motivation of the Simplex Method, Simplex Algorithm, Improving a Nonoptimal Basic Feasible Solution, Two Phases of the Simplex Method, Revised Simplex Method, Duality in Linear Programming.

Nonlinear Programming 10 Hours

Introduction, Unimodal Function, Elimination Methods: Unrestricted Search, Exhaustive Search, Dichotomous Search, Interval Halving Method Fibonacci Method, Golden Section Method, Comparison of Elimination Methods. Interpolation Methods: Quadratic Interpolation Method, Cubic Interpolation Method, Direct Root Methods. Direct Search Methods: Random Search Methods, Grid Search Method, Univariate Method, Pattern Directions, Powell’s Method.

Geometric Programming 10 Hours

Posynomial, Unconstrained Minimization Problem, Solution of an Unconstrained Geometric Programming Program Using Differential Calculus, Solution of an Unconstrained Geometric Programming Problem Using Arithmetic–Geometric Inequality, Primal–Dual Relationship and Sufficiency Conditions in the Unconstrained Case, Constrained Minimization, Solution of a Constrained Geometric Programming Problem, Primal and Dual Programs in the Case of Less-Than Inequalities, Geometric Programming with Mixed Inequality Constraints, Complementary Geometric Programming.

Stochastic Programming: Basic Concepts of Probability Theory, Stochastic Linear Programming, Stochastic Nonlinear Programming.

Modern Methods **10 Hours**

Genetic Algorithms, Simulated Annealing, Particle Swarm Optimization, Ant Colony Optimization, Optimization of Fuzzy Systems, Neural-Network-Based Optimization.

Text Book:

1. Singiresu S Rao, “Engineering Optimization Theory and Practice", John Wiley and sons, 4th Edition 2009.

Reference Book:

1. Edwin K. P. Chong and Stanislaw. Zak “An Introduction to Optimization”, John Wiley and sons, 2nd Edition 2001.

SDS242 COMPUTATIONAL LINGUISTICS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 Hours**

What is computational linguistics? Ambiguity and uncertainty in language, regular languages, and their limitations, finite-state automata, morphology.

Context Free Grammars **10 Hours**

Constituency, CFG definition, use and limitations. Chomsky Normal Form. Top-down parsing, bottom-up parsing, and the problems with each. The desirability of combining evidence from both directions

Programming in Python **10 Hours**

An introduction to programming from square one. Why Python? Variables, numbers, strings, arrays, dictionaries, conditionals, iteration. The NLTK (Natural Language Toolkit).

Word Sense Disambiguation and Clustering **10 Hours**

Homonomy, polysemy, different meanings, the power of context. Language neighbourhood as a vector. Agglomerative clustering. Clustering by expectation maximization. Using clustering to discover different word senses. Semi-supervised document classification.

Machine Translation **10 Hours**

Probabilistic models for machine translation system, alignment, translation, language generation. machine translation evaluation.

Text Book

1. Daniel Jurafsky and James H. SPEECH and LANGUAGE PROCESSING: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, Second Edition.

2. Chris Manning and Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

Reference Books:

1. Allen, James. 1995. – Natural Language Understanding. Benjamin/Cummings, 2ed.Bharathi, A Vineet Chaitanya and Rajeev Sangal. 1995.
2. Natural Language Processing- A Pananian Perspective. Prentice Hll India, Eastern Economy.
3. Eugene Cherniak: Statistical Language Learning, MIT Press, 1993.

SDS243 BIOINFORMATICS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction, Scope and Importance **8 Hours**

Important contributions, Aims and Tasks of Bioinformatics, Applications of Bioinformatics, Challenges and Opportunities, Introduction to NCBI data model, Various file formats for biological sequences, The Data: Storage and Retrieval , Basic Principles, The Data, Data Quality, Data Representation.

Bioinformatics Database **10 Hours**

Importance of Databases, Characteristics and Categories of Bioinformatics Database, Navigating Databases, Biological Databases, Primary Sequence Databases, Composite Sequence Databases, Secondary Databases, Nucleic Acid Sequence Databases, Structure Databases: File Formats, Protein Structure, PDB, MMDB, CATH, Other Database Enzyme, MEROPS, BRENDA, Pathway databases, Bibliographic Databases, Specialized Genomic Resources, Analysis Packages.

Sequence Align Methods **12 Hours**

Sequence Analysis of Biological Data, Significance of Sequence Alignment, Pairwise Sequence Alignment Methods, Use of Scoring Matrices and Gap Penalties in Sequence Alignments, Multiple Sequence Alignment Methods - Tools and Application of multiple sequence alignment, Gene Predictions Strategies, Protein Prediction Strategies, Phylogenetic Trees and Multiple Alignments.

Bioinformatics Algorithms **12 Hours**

Biological Algorithms versus Computer Algorithms, Exhaustive Search, Mapping Algorithms, Motif Finding Problem, Search Trees, Finding a Median String, Greedy Approach to Motif Finding, DNA Sequence comparison - Manhattan Tourist Problem - Edit Distance and Alignments - Longest Commons Subsequences - Global Sequence Alignment - Scoring Alignment - Local Sequence Alignment – Alignment with Gap Penalties - Multiple Alignment, DNA Sequencing, Shortest Superstring Problem, DNA arrays as an alternative sequencing techniques.

Biostatistics & Tools **10 Hours**

Handling Univariate and Bivariate Data, Measures of Central Tendency, Measures of Dispersion, Skewness & Kurtosis, Correlation and Regression.
Local Alignment Search Tool (BLAST), Purpose of BLAST, BLAST Analysis, Purpose of BLAST II, Scoring Metrics, PAM, BLOSUM, Working of BLAST, Introduction to HMMER.

Text Books:

1. Bioinformatics - Concepts, Skills, and Applications, S.C. Rastogi, Namita Mendiratta, Parag Rastogi, Second Edition, CBS Publishers, 2003.
2. An Introduction to Bioinformatics Algorithms, Neil C Jones and Pavel A Pevzner, MIT Press, 2004.

Reference Books:

1. Bioinformatics: Databases, Tools, And Algorithms., Orpita Bosu, Simminder Kaur Thukral , Oxford University Press Publisher, 2007.
2. Fundamentals of Mathematical Statistics., S.C. Gupta and V.K. Kapoor, Eleventh Edition, Sultan Chand & Sons Publishers, 2007.
3. Internet Resources

SDS251 CLOUD COMPUTING & VIRTUALIZATION

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 Hours**

Introduction, Cloud Infrastructure Cloud computing, Cloud computing delivery models and services, Ethical issues, Cloud vulnerabilities, Cloud computing at Amazon, Cloud computing the Google perspective, Microsoft Windows Azure and online services, Open-source software platforms for private clouds, Cloud storage diversity and vendor lock-in, Energy use and ecological impact, Service level agreements, User experience and software licensing. Exercises and problems.

Computing **10 Hours**

Cloud Computing: Application Paradigms. Challenges of cloud computing, Architectural styles of cloud computing, Workflows: Coordination of multiple activities, Coordination based on a state machine model: The Zookeeper, The Map Reduce programming model, A case study: The Grep The Web application , Cloud for science and engineering, High-performance computing on a cloud, Cloud computing for Biology research, Social computing, digital content and cloud computing.

Virtualization **10 Hours**

Cloud Resource Virtualization. Virtualization, Layering and virtualization, Virtual machine monitors, Virtual Machines, Performance and Security Isolation, Full virtualization and paravirtualization, Hardware support for virtualization, Case Study: Xen a VMM based paravirtualization, Optimization of network virtualization, vBlades, Performance comparison of virtual machines, The dark side of virtualization, Exercises and problems.

Management/Scheduling **10 Hours**

Cloud Resource Management and Scheduling. Policies and mechanisms for resource management, Application of control theory to task scheduling on a cloud, Stability of a two-level resource allocation architecture, Feedback control based on dynamic thresholds, Coordination of specialized autonomic performance managers, A utility-based model for cloud-based Web services, Resourcing bundling: Combinatorial auctions for cloud resources, Scheduling algorithms for computing clouds, Fair queuing, Start-time fair queuing, Borrowed virtual time, Cloud scheduling subject to deadlines, Scheduling Map Reduce applications subject to deadlines, Resource management and dynamic scaling, Exercises and problems.

Security **08 Hours**

Cloud Security, Cloud Application Development. Cloud security risks, Security: The top concern for cloud users, Privacy and privacy impact assessment, Trust, Operating system security, Virtual machine

Security, Security of virtualization, Security risks posed by shared images, Security risks posed by a management OS, A trusted virtual machine monitor, Amazon web services: EC2 instances, Connecting clients to cloud instances through firewalls, Security rules for application and transport layer protocols in EC2, How to launch an EC2 Linux instance and connect to it, How to use S3 in java, Cloud-based simulation of a distributed trust algorithm, A trust management service, A cloud service for adaptive data streaming, Cloud based optimal FPGA synthesis .Exercises and problems.

Text Book:

1. Dan C Marinescu: Cloud Computing Theory and Practice. Elsevier(MK) 2013.

Reference Books:

1. Rajkumar Buyya , James Broberg, Andrzej Goscinski: Cloud Computing Principles and Paradigms, Willey 2014.
2. John W Rittinghouse, James F Ransome:Cloud Computing Implementation, Management and Security, CRC Press 2013.

SDS152 WEB DATABASES & INFORMATION SYSTEMS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Web-based Information system **10 hours**

Web -Based Information Systems, Applications: electronic commerce, Variants of Web database access, Basic Web Standards, architecture. HTTP, Forms, Server-Side Programming and CGI Alternatives to CGI: Java Servlets and server APIs Browser detection, state, cookies and redirects

Using n-tiered architectures to implement secure and scalable systems **10 hours**

Web App Architectures: Multi-Tier (2-Tier, 3-Tier) Model-Viewer-Controller (MVC)

Database-driven websites and applications **10 Hours**

− Utilize JavaScript to improve database driven websites. − Critical components of the modern Web infrastructure: DNS, CDN, etc Critical components of the modern Web infrastructure: DNS, CDN, etc

DBMS and WWW: **10 hours**

Introduction Off-Line access to databases Static and Dynamic Web Pages SQL embedded in HTML CGI solution to database gateway Internet database connector JDBC: databases the Java way Solutions from database vendors Association rule mining

XML and its alternatives **10 Hours**

HTTP, XML, SQL, JavaScript, AJAX XML and its Alternatives: XML: Basics of XML, namespace, schema languages, XSLT and XPath, alternatives to XML, SQL, CSS, RSS, and others.

Text Books:

1. PROFESSIONAL WEB 2.0 PROGRAMMING: USING XHTML, CSS, JAVASCRIPT AND AJAX By Eric Van Der, Danny Ayers, Erik Bruchez
2. Weaving a Website - Programming in HTML, Javascript, Perl, and Java

Reference Book:

1. Mastering HTML, CSS & Javascript Web Publishing Paperback – 15 Jul 2016 by Laura Lemay (Author), Rafe Colburn (Author), Jennifer Kyrnin (Author)

SDS253 SOCIAL & INFORMATION NETWORK ANALYSIS

Total Teaching Hours: 50 **No. of Credits : 05**

Syllabus

Introduction **10 hours**

Overview: Aspects of Networks, Central Themes and Topics .
Graphs: Basic Definitions, Paths and Connectivity, Distance and Breadth-First Search, Network Datasets: An Overview

The Small-World Phenomenon **10 hours**

Six Degrees of Separation, Structure and Randomness, Decentralized Search, Empirical Analysis and Generalized Models, Core-Periphery Structures and Difficulties in Decentralized Search, Advanced Material: Analysis of Decentralized Search
Positive and Negative Relationships: Structural Balance , Characterizing the Structure of Balanced Networks, Applications of Structural Balance , A Weaker Form of Structural Balance , Advanced Material: Generalizing the Definition of Structural Balance.

Cascading Behaviour in Networks **10 hours**

Diffusion in Networks, Modeling Diffusion through a Network, Cascades and Clusters, Diffusion, Thresholds, and the Role of Weak Ties, Extensions of the Basic Cascade Model, Knowledge, Thresholds, and Collective Action, Advanced Material: The Cascade Capacity

Epidemics : Diseases and the Networks that Transmit Them , Branching Processes , The SIR Epidemic Model , The SIS Epidemic Model , Synchronization , Transient Contacts and the Dangers of Concurrency , Genealogy, Genetic Inheritance, and Mitochondrial Eve , Advanced Material: Analysis of Branching and Coalescent Processes

Power Laws and Rich-Get-Richer Phenomena **10 hours**

Popularity as a Network Phenomenon , Power Laws, Rich-Get-Richer Models , The Unpredictability of Rich-Get-Richer Effects , The Long Tail , The Effect of Search Tools and Recommendation Systems Advanced Material: Analysis of Rich-Get-Richer Processes.
The structure of the Web : The World Wide Web, Information Networks, Hypertext, and Associative Memory The Web as a Directed Graph , The Bow-Tie Structure of the Web, The Emergence of Web 2.0 U

Link Analysis and Web Search **10 hours**

Searching the Web: The Problem of Ranking , Link Analysis using Hubs and Authorities , PageRank, Applying Link Analysis in Modern Web Search , Applications beyond the Web, Advanced Material: Spectral Analysis, Random Walks, and Web Search

Strong and Weak Ties: Triadic Closure , The Strength of Weak Ties , Tie Strength and Network Structure in Large-Scale Data , Tie Strength, Social Media, and Passive Engagement Closure, Structural Holes, and Social Capital, Advanced Material: Betweenness Measures and Graph Partitioning

Text Book:

1. "Networks, Crowds, and Markets Reasoning about a Highly Connected World", David Easley, Cornell University, New York, Jon Kleinberg, Cornell University, New York, 2010

Reference Books:

1. "Networks: An Introduction by M. E. J. Newman, a college-level textbook about the science of networks.", M. E. J. Newman Hardback, Oxford University Press, 2010