

INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM



Curriculum and Syllabus for the PG Course e-MTech Programme In Artificial Intelligence, Machine Learning and Data Science

Contents

SEMESTER I	4
DSC511 Statistical Foundations for Data Science [2-0-0-2]	4
DSC512 Programming and Data Structures [2-0-2-3]	6
DSC513 Introduction to Data Science [2-0-2-3]	8
DSC522 Artificial Intelligence Engineering [2-0-2-3]	10
DSC523 Data Mining [3-0-0-3]	12
SEMESTER II	14
DSC521 Mathematical Foundations for Data Science [2-0-0-2]	14
DSC611 Machine Learning:Principles and Practices [3-0-0-3]	16
DSCXXX Designing MLoPs for Enterprises [3-0-0-3]	18
DSCXXX AI and ML for Big Data [1-0-0-1]	20
DSC613 Big Data Analytics [2-0-2-3]	22
SEMESTER III	25
DSC612 Neural Networks and Deep Learning [3-0-0-3]	25
DSC624 Natural Language Processing and Large Language Models [2-0-2-3]	27

e-M.Tech in Artificial Intelligence, Machine Learning and Data Science

General Course Structure

Semester- I					
Course Code	Course Name	L	T	P	C
DSC511	Statistical Foundations for Data Science	2	0	0	2
DSC512	Programming and Data Structures	2	0	2	3
DSC513	Introduction to Data Science	2	0	2	3
DSC522	Artificial Intelligence Engineering	2	0	2	3
DSC523	Data Mining	3	0	0	3
Semester- II					
DSC521	Mathematical Foundations for Data Science	2	0	0	2
DSC611	Machine Learning:Principles and Practices	3	0	0	3
DSCXXX	Designing MLoPs for Enterprises	3	0	0	3
DSCXXX	AI and ML for Big Data	1	0	0	1
DSC613	Big Data Analytics	2	0	2	3
Semester- III					
DSC612	Neural Networks and Deep Learning	3	0	0	3
DSC624	Natural Language Processing and Large Language Models	2	0	2	3
CSE711	Project(Phase I)				14
Semester- IV					
CSE721	Project (Phase II)				14
Total Credits					60

Electives:

- AI and ML for Big Data
- Information Retrieval
- Ethics for Data Science
- Advanced Topics in Data Processing
- Neural Networks and Deep Learning
- Big Data Analytics
- Data Visualization and Predictive analytics
- Designing MLoPs for Enterprises

L	T	P	C
2	0	0	2

SEMESTER I

DSC511 Statistical Foundations for Data Science

Course Objectives

- To learn basic and some advanced concepts in probability and statistics.
- To learn the concepts of statistics and random process in solving problems arising in data science.

Course Outcomes

- Students will be able to model uncertain phenomena using probability models and calculate the uncertainty in systems where such phenomena are a part of the system.
- Students will be able to implement statistical analysis techniques for solving practical problems.

Syllabus

Probability: Sample space, events and axioms; conditional probability; Bayes theorem; Random variables; Standard discrete and continuous probability distributions; Expectations and moments; Covariance and correlation; Linear Regression; Central limit theorem.

Statistics: Sampling distributions of the sample mean and the sample variance for a normal population; Point and interval estimation; Sampling distributions (Chi-square, t,F,Z), Hypothesis testing ; One tailed and two-tailed tests; Analysis of variance, ANOVA, One way and two way classifications.

Random Processes: Definition and classification of random processes, Poisson process, Gaussian white noise. Statistical analysis using R.

Learning Resources

1. S. Ross, Introduction to Probability and Statistics for and Engineers and Scientists, Third Edition, Elsevier, 2004.
2. G. R. Grimmett and D. R. Stirzaker, Probability and Random Processes, Oxford University Press, 2001
3. R.V. Hogg, J.W. McKean & A. Craig, Introduction to Mathematical Statistics, 6th Edition.

4. Montgomery, D. C. and G. C. Runger, Applied Statistics and Probability for Engineers. 5th Edition. John Wiley & Sons, Inc., NY, USA. 2009
5. Robert H. Shumway and David S. Stoffer, Time Series Analysis and Its Applications with R Examples, Third edition, Springer Texts in Statistics, 2006.
6. Athanasios Papoulis, Probability Random Variables and Stochastic Processes, 4th edition, McGraw-Hill, 2002.

L	T	P	C
2	0	2	3

DSC512 Programming and Data Structures

Course Objectives

The course is intended to provide the foundations of the practical implementation and usage of Algorithms and Data Structures.

- Ensure that the student evolves into a competent programmer capable of designing and analysing implementations of algorithms and data structures for different kinds of problems.
- Expose the student to the algorithm analysis techniques, to the theory of reductions, and to the classification of problems into complexity classes like NP.

Course Outcomes

- Design and analyse programming problem statements.
- Choose appropriate data structures and algorithms, understand the ADT/libraries, and use it to design algorithms for a specific problem.
- Understand the necessary mathematical abstraction to solve problems.
- Come up with analysis of efficiency and proofs of correctness
- Comprehend and select algorithm design approaches in a problem specific manner.

Syllabus

Introduction: Introduction to Data Structures and Algorithms, Review of Basic Concepts, Asymptotic Analysis of Recurrences. Randomized Algorithms. Randomized Quicksort, Analysis of Hashing algorithms.

Algorithm Analysis Techniques: Amortized Analysis, Application to Splay Trees, External Memory ADT - B-Trees, Priority Queues and Their Extensions: Binomial heaps, Fibonacci heaps, applications to Shortest Path Algorithms, Partition ADT: Weighted union, path compression, Applications to MST. Algorithm Analysis and Design Techniques.

Dynamic Programming: Greedy Algorithms Bellman-Ford. Network Flows-Max flow, min-cut theorem, Ford-Fulkerson, Edmonds-Karp algorithm.

Intractable Problems: Polynomial Time, class P, Polynomial Time Verifiable Algorithms, class NP, NP completeness and reducibility, NP Hard Problems, Approximation Algorithms.

Learning Resources

1. Introduction to Algorithms, by T. H. Cormen, C. E. Lieserson, R. L. Rivest, and C. Stein, Third Edition, MIT Press.
2. Fundamentals of Data Structures in C by Horowitz, Sahni, and Anderson-Freed, Universities Press
3. Algorithms, by S. Dasgupta, C. Papadimitrou, U Vazirani, Mc Graw Hill.
4. Algorithm Design, by J. Klienbergs and E. Tardos, Pearson Education Limited.

L	T	P	C
2	0	2	3

DSC513 Introduction to Data Science

Course Objectives

- To understand the basic concepts of Data science
- Ability to apply Data Science in different domains
- Do exploratory analysis on a given data
- Responsible Practices on Data Science

Course Outcomes

- Use R to carry out basic statistical modelling and analysis
- Explain the significance of exploratory data analysis (EDA) in data science. Apply basic tools (plots, graphs, summary statistics) to carry out EDA
- Apply EDA and the Data Science process in a case study.
- Create effective visualization of given data
- Describe the Data Science Process and how its components interact.
- Work effectively in teams on data science projects.
- Incorporating responsible practices in Data Science projects.

Syllabus

Introduction: Data science, data analytics, machine learning, and Artificial Intelligence. AI and Data Science in your company, AI and society. Role of Data.

Data Science Programming: Introduction to R, R packages, R Markdown, Programming e.g. functions, loops, if/else, comments, Tidy data, Tabular data and data import, Strings and regular expressions. Familiarization of functions in TidyR package.

Manipulation of Data: Data Wrangling, Data manipulation dplyr. Plotting- Visualization with ggplot2. Statistical inference using R, What-if analysis, case studies. Qualitative Data Analysis, Exploring Maps in R programming

Application: Exploratory Data Analysis and the Data Science Process, Basic tools (plots, graphs and summary statistics) of EDA. Statistical Modelling, Missing Values. Case studies.

Responsible Data Science using R: Introduction to Responsible Data Science: Concepts

of fairness, accountability, transparency, and privacy. Practical Implementation: Bias detection and mitigation techniques using R. Model Explainability and Interpretability. Privacy-preserving techniques using R.

Case studies demonstrating responsible data handling and ethical considerations using R. Discussions on privacy, security, ethics.

Learning Resources

1. Wickham, Hadley, and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.", 2016.
2. Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from The Frontline. O'Reilly. 2014.
3. Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. ISBN 1449361323. 2013.
4. Martin Braschler, Thilo Stadelmann, Kurt Stockinger Applied Data Science Lessons Learned for the Data-Driven Business, Springer 2019.
5. Peter Bruce and Andrew Bruce, Practical Statistics for Data Scientists, Published by O'Reilly Media 2017.
6. Software for Data Analysis: Programming with R (Statistics and Computing), John M. Chambers, Springer.
7. Joshua Kroll et al., "Accountable Algorithms," University of Pennsylvania Law Review, 2017.
8. Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness and Machine Learning. Fairmlbook.org, 2019.
9. Max Kuhn and Kjell Johnson. Applied Predictive Modeling. Springer, 2013. (Includes fairness metrics and techniques using R)
10. Biecek, Przemyslaw, and Staniak, Mateusz. "Explanatory Model Analysis: Explore, Explain, and Debug Predictive Models with DALEX." CRC Press, 2021.

L	T	P	C
2	0	2	3

DSC522 Artificial Intelligence Engineering

Course Objectives

- Cover various paradigms that come under the broad umbrella of AI.

Course Outcomes

- The students are expected to have the ability to develop an understanding of where and how AI can be used.

Syllabus

Introduction: AI:, Brief history. Agents and rationality, task environments, agent architecture types. Search and Knowledge representation. Search spaces Uninformed and informed search.

Techniques: Hill climbing, simulated annealing, genetic algorithms. Logic based representations (PL, FoL(Syntax and semantics , Using first order logic, Inference in first order logic.)) and inference Prolog. Rule based representations, forward and backward chaining, matching algorithms. Probabilistic reasoning and uncertainty.

Learning: Uncertainty and methods to handle it. Learning. Forms of learning. Statistical methods: Bayesian learning (Learning Bayes net structures), AI Modelling Techniques.

Building Intelligent Systems: From Model to AI-enabled System, Goals and Success Measures for AI-Enabled Systems, Tradeoffs among Modeling Techniques, Dealing with Mistakes, Software Architecture of AI-Enabled Systems, Deployment and Process, Introduction to Domain Specific AI, Building Fair AI, Explainability and Interpretability, Introduction to Human-AI Interaction.

Applications: Applications to NLP, vision, robotics.

Learning Resources

1. Russel,S., and Norvig,P., (2015), Artificial Intelligence: A Modern Approach, 3rd Edition, Prentice Hall
2. Lang, Q. (1997), Intelligent Planning: A decomposition and abstraction-based approach, Springer Verlag, Berlin Heidelberg.
3. Patterson, Introduction to AI and Expert Systems, 3rd edition PHI.

4. Bosch, Jan, Helena Holmström Olsson, and Ivica Crnkovic. "Engineering AI systems: A research agenda." Artificial Intelligence Paradigms for Smart Cyber-Physical Systems. IGI global, 2021. 1-19.

L	T	P	C
3	0	0	3

DSC523 Data Mining

Course Objectives

- To introduce basic concepts, tasks, methods, and techniques in data mining.
- Examine the types of the data to be mined and apply pre-processing methods on raw data
- Apply various data mining problems and their solutions

Course Outcomes

At the end of the course the students will able to:

- Develop an understanding of the data mining process and issues.
- Understand various techniques for data mining
- Apply the techniques in solving data mining problems using data mining tools and systems
- Expose various real-world data mining applications

Syllabus

Data Mining Concepts: Introduction to modern data analysis (Data visualization; probability; histograms; multinomial distributions), Data Mining and Knowledge Discovery in Databases, Data Mining Functionalities, Data Pre-processing, Data Cleaning, Data Integration and Transformation, Data Reduction, Data Discretization and Concept Hierarchy Generation, examples.

Data mining algorithms: Association Rule Mining, Classification and Prediction: Issues Regarding Classification and Prediction, Classification by Decision Tree Regression, Bayesian Classification, Rule Based Classification, Classification by Back propagation, Support Vector Machines, Associative Classification, Lazy Learners, Random Forest, Other Classification Methods.

Cluster Analysis: Types of Data in Cluster Analysis, Model-Based Clustering Methods, Hierarchical and Partitioning methods. Outlier Analysis.

Applications and trends in Data Mining: Sequential Pattern Mining; Mining Text and Web data, Mining Spatiotemporal and Trajectory Patterns, Multivariate Time Series (MVTs) Mining.

Programming Assignments

- Basics of R/python
- Data Pre-processing and cleaning
- Data Reduction
- Association rule mining
- Classification and Prediction
- Cluster Analysis
- Outlier analysis
- Mining text and web
- Experiment based on applications of data mining

Learning Resources

1. Alex Berson, Stephen J. Smith, "Data Warehousing, Data Mining, & OLAP", Tata Mcgraw-Hill, 2004.
2. Jiawei Han. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers
3. Anahory and Murray ., Data warehousing in the real world , Pearson Education / Addison Wesley.
4. Berry Micheal and Gordon Linoff, Mastering Data Mining. John Wiley & Sons Inc.
5. Margaret H. Dunham Data Mining: Introductory and Advanced Topics. Prentice Hall
6. Hadzic F., Tan H. & Dillon T.S. "Mining data with Complex Structures" Springer, 2011
7. Yates R. B. and Neto B. R. "Modern Information Retrieval" Pearson Education, 2005
8. Tan P. N., Steinbach M & Kumar V. "Introduction to Data Mining" Pearson Education, 2006
9. Christopher D.M., Prabhakar R. & Hinrich S. "Introduction to Information Retrieval" Cambridge UP Online edition, 2009
10. Witten, E. Frank, M. Hall. "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann Publishers, 2011.

L	T	P	C
2	0	0	2

SEMESTER II

DSC521 Mathematical Foundations for Data Science

Course Objectives

- To learn important linear algebra techniques and their applications in data mining, machine learning, pattern recognition.
- To explain the basic mathematical concepts of optimization
- To provide the skills necessary to solve and interpret optimization problems in engineering.

Course Outcomes

- Students will be able to model problems through abstract structures and arrive at insights or solutions by manipulating these models using their properties.
- Understand the different methods of optimization and be able to suggest a technique for a specific problem.

Syllabus

Linear Algebra: Matrices, Vectors and their properties (determinants, traces, rank, nullity, etc.); Inner products; Distance measures; Projections; Notion of hyper planes; half-planes; Positive definite matrices; Eigenvalues and eigenvectors.

Numerical Linear Algebra: System of linear equations; Matrix factorizations; QR Decomposition; Singular value decompositions; Cholesky Factorization; Least squares Problem; Finding roots of an equation: Newton Raphson method.

Optimization: Unconstrained optimization; Necessary and sufficiency conditions for optima; Gradient descent methods; constrained optimization, convex sets, KKT conditions.

Learning Resources

1. Matrix Computations by Gene H. Golub, C.F. Van Loan, The Johns Hopkins University Press.
2. G. Strang , Introduction to Linear Algebra, Wellesley-Cambridge Press, Fifth edition, USA,2016.
3. Numerical Linear Algebra by Lloyd N. Trefethen and David Bau, III, SIAM, Philadelphia,1997.

4. D. S. Watkins, Fundamentals of Matrix Computations, 2nd Edition, John Wiley & Sons, 2002.
5. David G. Luenberger, Optimization by Vector Space Methods, John Wiley & Sons (NY), 1969.
6. An introduction to optimization by Edwin K. P. Chong and Stanislaw H. Zak, 4th Edition, Wiley, 2013.

L	T	P	C
3	0	0	3

DSC611 Machine Learning: Principles and Practices

Pre-requisites

Basic programming knowledge, Probability and statistics

Course Objectives

- To provide basis understandings on mathematical foundations and concepts of machine learning
- To provide an in-depth introduction to supervised, unsupervised and reinforcement learning algorithms.
- To design and implement machine learning solutions to classification, regression, and clustering problems.

Course Outcomes

- Develop an appreciation for what is involved in learning from data.
- Understand a wide variety of learning algorithms.
- Understand how to apply a variety of learning algorithms to appropriate data.
- Understand how to perform evaluation of learning algorithms and model selection.
- Apply machine learning methods to real word problems

Syllabus

Basic Principles: Introduction, Computational Learning Theory (CLT): PAC learning, Sample complexity, VC-dimension, Bias and variance, Experimental Evaluation: overfitting and underfitting, Cross-Validation, cost function optimization. Bagging, boosting.

Supervised Learning: Review of Linear algebra and convex optimization, Gradient descent based optimization: Batch and stochastic gradient descent. Regression algorithms: Simple linear regression, multiple linear regression, polynomial regression, L1 and L2 Regularization.. Logistic Regression, Gaussian discriminant analysis (Naïve bayes, Naïve bayes with Laplace smoothing), Binary and multiclass Classification: SVM (Quadratic programming solution to finding maximum margin separators. Kernels for learning non-linear functions, Kernel Optimization), model selection. Multiclass Classification: Generalization bounds, Multiclass SVM.

Unsupervised Learning: Clustering (Spectral clustering learning), Expectation Maximization, Mixture of Gaussians, Hidden Markov Models.

Probabilistic Models, Kernel Methods and Latent Space Models: Probabilistic Models Maximum Likelihood Estimation, MAP, Probabilistic Principal Component Analysis, Latent Dirichlet allocation, Kernel Methods: Basics, Gaussian Processes, Kernels on Strings, trees, graphs, Latent Space Models: Independent Component Analysis.

Recent trends in ML: Federated learning: Concepts, architecture and algorithms, Horizontal and vertical federated learning, federated transfer learning, distributed machine learning.

Programming Assignments

- Experiments based on validation of models, regression and classification.
- Experiments on Gradient descent and stochastic gradient descent.
- Case studies using SVN and Multiclass SVM
- Experiments on Back propagation
- Implementations of Dimension reduction, EM algorithm and HMM
- Implementation of MAP, PCA, LDA, Kernel methods.
- Apply of machine learning methods to real word applications (Case studies)
- Experiments on building federated learning models

Learning Resources

1. Tom Mitchell. Machine Learning. McGraw Hill, 1997.
2. Machine Learning: A Probabilistic Perspective, Kevin P Murphy, MIT Press.
3. Christopher M. Bishop. Pattern Recognition and Machine Learning. Springer 2006.
4. Richard O. Duda, Peter E. Hart, David G. Stork. Pattern Classification. John Wiley & Sons, 2006.
5. Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer 2009.
6. MacKay, David. Information Theory, Inference, and Learning Algorithms. Cambridge, UK: Cambridge University Press, 2003.
7. Yang, Qiang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. "Federated learning." Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan & Claypool publishers, 2019.

L	T	P	C
3	0	0	3

DSC XXX Designing MLOPs for Enterprises

Course Objectives

- Ability to design end-to-end machine learning systems for practical problems
- Ability to identify key metrics to optimize model performance
- Ability to critically evaluate various deployment options

Course Outcomes

At the end of this course, students will be able to:

- Describe and apply learning algorithms to datasets of massive size
- Develop optimized models for learning tasks like classification and clustering on large datasets
- Analyse the feasibility of scalable learning algorithms

Syllabus

Introduction to DevOps: SDLC, Virtualization: Containers, Container Orchestration Systems, Cloud platforms, CI/CD: Continuous Integration – Configuration Management, Deployment and Delivery phases, Continuous monitoring, Continuous Testing.

Basic Concepts: Evolution of MLOps, Data-centric AI, ML Development Lifecycle, MLOps Approach, Features of MLOps, ML Data Lifecycle in Production, MLOps maturity levels, ML artifacts, MLOps workflows.

Machine Learning Pipelines and automation: CI/CD for Machine Learning, ML model serving, Data pipelines, Data drift, ML pipelines: Data ingestion, Feature engineering, Hyperparameter optimization, testing and packaging. Model management, Model deployment and monitoring, feedback, orchestration pipelines for ML workflows, ML security, Real-time Streaming ML models, Deployment on edge devices, Automated ML, case studies.

Learning Resources

1. Treveil, Mark, Nicolas Omont, Clément Stenac, Kenji Lefevre, Du Phan, Joachim Zentici, Adrien Lavoillotte, Makoto Miyazaki, and Lynn Heidmann. Introducing MLOps. O'Reilly Media, 2020.
2. Burkov, Andriy. Machine Learning Engineering. True Positive Inc. , 2020.
3. Ameisen, Emmanuel. Building Machine Learning Powered Applications. O'Reilly Media, 2020.

4. Alla, Sridhar, and Suman Kalyan Adari. Beginning MLOps with MLFlow. Apress, 2021.
5. Rao, Dattaraj. Keras to Kubernetes: The Journey of a Machine Learning Model to Production. John Wiley & Sons, 2019.
6. Sculley, David, et al. "Machine learning: The high interest credit card of technical debt." (2014).
7. Jez Humble, David Farley. Continuous Delivery.

L	T	P	C
1	0	0	1

DSCXXX AI and ML for Big Data

Course Objectives

- To introduce the concepts, fundamentals and methodologies for generating models from Big Data.
- To impart knowledge on building scalable learning models.
- To provide an understanding of frameworks to implement scalable algorithms for the analysis of massive data.

Course Outcomes

At the end of this course, students will be able to:

- Describe and apply learning algorithms to datasets of massive size
- Develop optimized models for learning tasks like classification and clustering on large datasets
- Analyse the feasibility of scalable learning algorithms

Syllabus

Introduction: Cross Industry Standard Process for Data Mining (CRISP-DM), Applications of Big data and ML.

Learning on Big Data: Frequent itemset mining for big data using PCY algorithm, Big Data Clustering and Classification, Representation Learning, Modelling, Clustering in non-Euclidean spaces, BFR Clustering Algorithm, BigCLAM for community detection, Local Sensitive Hashing, CURE, Recommendation Systems – Content Based & Collaborative Filtering, CUR Decomposition for dimensionality reduction, scaling ML algorithms.

AI and ML solutions for Big Data: Requirements and Applications for Smart Healthcare, Digital Learning, Customer Experience Management, Business Intelligence , Electric Power Systems.

Learning Resources

1. Jure Leskovec, Anand Rajaraman, Jeffrey D. Ullman, “Mining of Massive Datasets”, Stanford School of Engineering, 3rd edition, 2020.
2. Anand Deshpande and Manish Kumar, “Artificial Intelligence for Big Data”, Packt publishers, 2018.

3. Shan Suthaharan, “Machine Learning Models and Algorithms for Big Data Classification”, Springer, 2015.
4. Daniele Apiletti, Elena Baralis, Tania Cerquitelli, Paolo Garza, Fabio Pulvirenti, Luca Venturini, “Frequent Itemsets Mining for Big Data: A Comparative Analysis”, Elsevier, 2017.
5. Nick Pentreath, “ Machine Learning with Spark”, Packt, 2014.
6. Md. Rezaul Karim, Shridhar Alla, “Scala and Spark for Big Data Analytics”, Packt, 2017.
7. Miltiades Lytras, Akila Sarirete, Anna Visvizi, Kwok Tai Chui, “AI and Big data analytics for Smart Healthcare”, Elsevier Academic Press, 2021.
8. Fernando Iafrate, “Artificial Intelligence and Big Data”, Wiley, 2018.
9. Moses Strydom, Sheryl Buckley, “AI and Big Data’s for Disruptive Innovation”, IGI Global, 2019.

L	T	P	C
2	0	2	3

DSC613 Big Data Analytics

Pre-requisites

Basic programming knowledge (python) and basics of statistics.

Course Objectives

- To familiarize students with big data analysis as a tool for analysing large complex dataset.
- To learn to use various techniques for mining data stream.
- Understand the applications using Map Reduce Concepts
- Provide hands on Hadoop Eco System
- To introduce programming tools PIG & HIVE in Hadoop echo system

Course Outcomes

At the end of the course the students will be able to:

- Process data in Big Data platform and explore the big data analytics techniques for business applications
- Analyse Map Reduce technologies in big data analytics
- Develop Big Data solutions using Hadoop Eco System
- Design efficient algorithms for stream data mining on big data platform

Syllabus

NoSQL Database: Schema less Models, Increasing Flexibility for Data Manipulation-Key Value Stores, Document Stores, Tabular Stores, Object Data Stores - Graph Databases, Big data for twitter, Big data for E-Commerce blogs.

Data visualization: Basics of data visualization, Principles used for visualization, Lie factor, Category based visualization, Visualization tools and techniques, Outlier detection using visualization, Visualization based data analysis techniques.

Big Data: Evolution of Big data, Best Practices for Big data Analytics - Big data characteristics - Big Data Use Cases, Characteristics of Big Data Applications, Big Data Modelling, HDFS performance and tuning, Map reduce algorithm, Hadoop Eco system Pig : Introduction to Pig, Execution Modes of Pig, Grunt, Pig Latin, User Defined Functions, Data Processing operators. Hive : Hive Shell, Hive Services, HiveQL, Tables, Querying Data and User Defined

Functions. Hbase : HBasics, Concepts, Clients, Example, Spark.

Mining Data Streams: Introduction to Streams Concepts, Stream Data Model and Architecture - Sampling Data in a Stream, Filtering Streams, Counting Distinct Elements in a Stream –Real time Analytics Platform (RTAP) applications, Case Studies, Real Time Sentiment Analysis- Stock Market Predictions.

Programming Assignments

Exploring data analysis techniques library of python

- Experiments of various data plotting and visualization techniques
- Programming on Hadoop
- Programming on PIG, HIVE, and Spark
- Implementation of Machine Learning techniques on Big Data
- Implementation of Stream data mining techniques

Learning Resources

1. Jure Leskovec, Anand Rajaraman and Jeffrey David Ullman, "Mining of Massive Datasets", Cambridge University Press, 2014.
2. Tom White , Hadoop: The Definitive Guide, 4th edition O'Reily Publications, 2015
3. Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman, "Big data for dummies" A wiley brand publications.
4. Holden Harau, "Learning Spark: Lightning-Fast Big Data Analysis", O-Reilly Publications
5. David Loshin, "Big Data Analytics: From Strategic Planning to Enterprise Integration with Tools, Techniques, NoSQL, and Graph", 2013.
6. Bart Baesens, "Analytics in a Big Data World: The Essential Guide to Data Science and its Applications", Wiley Publishers, 2015.
7. Kim H. Pries and Robert Dunnigan, "Big Data Analytics: A Practical Guide for Managers " CRC Press, 2015.
8. EMC Education Services, "Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data", Wiley publishers, 2015.
9. Dietmar Jannach, Markus Zanker, Alexander Felfernig and Gerhard Friedrich "Recommender Systems: An Introduction", Cambridge University Press, 2010.

10. Jimmy Lin, Chris Dyer and Graeme Hirst, "Data-Intensive Text Processing with MapReduce", Synthesis Lectures on Human Language Technologies, Vol. 3, No. 1, Pages 1-177, Morgan Claypool publishers, 2010.

L	T	P	C
3	0	0	3

SEMESTER III

DSC612 Neural Networks and Deep Learning

Course Objectives

- To provide an in-depth introduction to neural network architectures and training procedures and its applications
- Introduce major deep learning algorithms, the problem settings, and their applications to solve real world problems.

Course Outcomes

- Able to implement neural network architectures and training procedures.
- Thoroughly Understanding the fundamentals of Deep Learning.
- Gaining knowledge of the different modalities of Deep learning currently used.
- Gaining knowledge about state-of the art models and Other Important Works in recent years.
- Learning the skills to develop Deep Learning based AI Systems (Use of Multiple packages etc.)

Syllabus

Neural Networks and its variants: Neural Networks and its variants, Multi-layer Perceptron, the neural viewpoint, Training Neural Network: Risk minimization, loss function, regularization, Neural Network model selection, and optimization.

Deep Learning Techniques: Deep Feed Forward network, regularizations, training deep models, dropouts, Convolutional Neural Network and Architectures, Deep Learning Hardware and Software. Recurrent Neural Network, Deep Belief Network, Autoencoders, Reinforcement learning – Passive and active, Generalization in RL, Policy Search, Deep Reinforcement Learning.

Tools and advanced techniques in neural network: Intro to Deep Learning Tools (Pytorch, Tensorflow, Caffe, Theano) CNN Architectures, Generative Models.

Application to Computer Vision: Computer vision overview, Historical context, Image Classification: Linear classification for Image I: Loss Functions and Optimization, Linear classification for Image II, Higher-level representations, image features, Softmax classifier, Object

Detection and Segmentation, Visualizing and Understanding.

Latest Trends: Generative Deep Learning models, Attention mechanisms and transformers, Self-supervised learning and contrastive learning.

Learning Resources

1. Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "Deep Learning"
2. Michael Nielson, "Neural Networks and Deep Learning"
3. Bishop, C. ,M., Pattern Recognition and Machine Learning, Springer, 2006
4. Zhang, Aston, et al. "Dive into deep learning." arXiv preprint arXiv:2106.11342 (2021).
5. Foster, David. Generative deep learning: teaching machines to paint, write, compose, and play. O'Reilly Media, 2019.

L	T	P	C
2	0	2	3

DSC624 Natural Language Processing and Large Language Models

Course Objectives

- Understand foundational concepts in natural language processing methods and strategies
- Evaluate the strengths and weaknesses of various NLP technologies and frameworks
- Implement Neural Language Models and Neural Machine Translation.
- Apply Transformers and Large Language models for building various NLP applications

Course Outcomes

- Analyse NLP tasks by applying fundamental techniques
- Implement neural network architectures for NLP tasks, including RNNs and attention mechanisms.
- Design and implement advanced NLP models utilising Transformer architectures and large language models
- Apply advanced techniques such as prompting and in-context learning to address real-world NLP challenges effectively.

Syllabus

Introduction to NLP: NLP Tasks, Challenges, Basics of Text processing, Tokenization, Normalisation, Vector representation of words: TF-IDF, Word2vec, Semantic similarity of word vectors, Machine Learning algorithms for Text classification and Sentiment analysis.

Probability and Language Model: Probability and NLP, Language Models, Markov Property, N-grams, Evaluating language models, Smoothing.

Neural Language Models: Neural networks for NLP tasks, Limitations of Feed Forward neural networks and CNN for NLP, Recurrent Neural Networks, Vanishing and Exploding gradients, LSTM, GRU, Encoder-Decoder models-Neural Machine Translation, Attention mechanism.

Transformers and Large Language Models: Transformers-self attention, Masked attention, Transformer Blocks, Language models with Transformers, Pre-training and Fine tuning, BERT and GPT models, Prompting, Retrieval augmented generation, Responsible and Ethical LLMs.

Programming Assignments

- Basics of text processing: Processing Raw Text
- Categorizing and Tagging words
- Text classification
- Sentiment Analysis
- Lexical Semantics
- Word embedding
- RNN for Text classification
- Neural Machine Translation
- Huggingface pipelines for LLM applications
- Exploring Langchain for LLM applications

Learning Resources

1. Jurafsky, Daniel, and James H. Martin. "Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.", 3rd Edition, 2024.
2. Bird, Steven, Ewan Klein, and Edward Loper. "NLTK tutorial: Introduction to natural language processing." Creative Commons Attribution 1037 (2005).
3. Dahl, Deborah Anna. "Natural language understanding with Python: combine natural language technology, deep learning, and large language models to create human-like language comprehension in computer systems." , Packt Publishing, (2023).
4. Amaratunga, Thimira. "Understanding large language models: learning their underlying concepts and technologies." , Apress Media, 2023.
5. Manning, Christopher, and Hinrich Schutze. Foundations of statistical natural language processing. , MIT press, 1999.