INDIAN INSTITUTE OF INFORMATION TECHNOLOGY KOTTAYAM

Curriculum and Syllabus for the PG Course

M.Tech Programme

In Computer Science and Engineering with Specialization in Bigdata & Machine Learning

For Working Professionals

CURRICULUM

M. Tech in Computer Science and Engineering with Specialization in Bigdata & Machine Learning

General Course Structure

Subject		L-T-P	Credits
Semester I			
DSC 511	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
BML511	Databases for Big-Data Applications	1-0-0	1
Semester II			
DSC 521	Mathematical Foundations for Data Science	2-0-0	2
DSC 523	Data Mining	3-0-0	3
BML521	Distributed Systems for Big Data Management and Processing	1-0-0	1
BML522	Big Data Visualization	1-0-0	1
Semester III			
DSC XXX	Elective I	3-0-0	3
DSC XXX	Elective II	3-0-0	3
DSC613	Big Data Analytics	2-0-2	3
BML XXX	Elective III	1-0-0	1
Semester IV			
DSC XXX	Elective IV	2-0-2	3
BML XXX	Elective V	2-0-0	2
BML XXX	Elective VI	1-0-0	1
Semester V			
CSE711	Project (Phase I)		14
Semester VI			
CSE721	Project (Phase II)		14
Total Credits		**60**	

Electives:

- Real time-Analytics
- Information Retrieval
- Ethics for Data Science
- Natural Language Processing
- Advanced Topics in Data Processing
- Designing MLOps for enterprises

- Neural Networks and Deep Learning
- Big Data Analytics
- Data Visualization and Predictive analytics
- Graph Algorithms and Mining
- Cloud Computing for Big Data

M. Tech in Computer Science and Engineering with Specialization in Bigdata & Machine Learning

Detailed Course Structure

Subject		L-T-P	Credits
Semester I			
DSC 511	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
BML511	Databases for Big-Data Applications	1-0-0	1
Semester II			
DSC 521	Mathematical Foundations for Data Science	2-0-0	2
DSC 523	Data Mining	3-0-0	3
BML521	Distributed Systems for Big Data Management and Processing	1-0-0	1
BML522	Big Data Visualization	1-0-0	1
Semester III			
DSC 611	Machine Learning: Principles and Practice	3-0-0	3
DSC 612	Neural Networks and Deep Learning	3-0-0	3
DSC 613	Big Data Analytics	2-0-2	3
BML611	Big Data Security	1-0-0	1
Semester IV			
DSC624	Natural Language Processing and Large Language Models	2-0-2	3

BML622	Designing MLOps for enterprises.	2-0-0	2
BML624	Realtime Big Data Analytics	1-0-0	1
Semester V			
CSE711	Project (Phase I)		14
Semester VI			
CSE721	Project (Phase II)		14
Total Credits		**60**	

CURRICULUM

SEMESTER I

DSC 511 Statistical Foundations for Data Science [2-0-0-2]

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.

DSC512 Programming and Data Structures [2-0-2-3]

Course Objectives

The course is intended to provide the foundations of the practical implementation and usage of Algorithms and Data Structures. One objective is to 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. The second objective is to 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. Klienberg and E. Tardos, Pearson Education Limited.

DSC513 Introduction to Data Science [2-0-2-3]

Course Objectives

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

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.

Syllabus

Introduction: data science, data analytics, machine learning, and Artificial Intelligence. AI 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. Web scraping, Text data and Natural Language Processing. Data Visualization, Data Science and Ethical Issues, Discussions on privacy, security, ethics.

Learning Resources

1. Cathy O'Neil and Rachel Schutt. Doing Data Science, Straight Talk from The Frontline. O'Reilly. 2014.
2. Foster Provost and Tom Fawcett. Data Science for Business: What You Need to Know about Data Mining and Data-analytic Thinking. ISBN 1449361323. 2013.
3. Martin Braschler,Thilo Stadelmann, Kurt Stockinger Applied Data Science Lessons Learned for the Data-Driven Business, Springer 2019.
4. Peter Bruce and Andrew Bruce,Practical Statistics for Data Scientists, Published by O'Reilly Media 2017.
5. An Introduction to Statistical Learning: with Applications in R, G James, D. Witten, T Hastie, and R. Tibshirani, Springer, 2013
6. Software for Data Analysis: Programming with R (Statistics and Computing), John M. Chambers, Springer.

BML511 Databases for Big-Data Applications [1-0-0]

Course Objectives

- This course will explore the origins of modern databases and the characteristics that distinguish them from traditional relational database management systems.
- Core concepts of NoSQL databases will be presented, followed by an exploration of how different database technologies implement these core concepts.
- Understand the impact of the cluster on database design
- Apply Nosql development tools on different types of NoSQL Databases
- To provides an in-depth understanding of terminologies and the core concepts behind big data problems.

Course Outcomes

Upon completion of this course, learners should be able to:

1. Demonstrate an understanding of the detailed architecture, define objects, load data, query data and performance tune Column-oriented NoSQL databases.
2. Explain the detailed architecture, define objects, load data, query data and performance tune Document-oriented NoSQL databases.
3. Evaluate NoSQL database development tools and programming languages.
4. Perform hands-on NoSql database lab assignments that will allow students to use the four NoSQL database types via products such as Cassandra, Hadoop Hbase, MongoDB and Neo4J

Prerequisite: Database Management, Basic programming knowledge (python), SQL and basics of statistics.

Syllabus

SQL & Big data Revolutions:
Introduction-Structured Query Language (SQL)- Data Types and Constraints in MySQL- SQL for Data Definition -SQL for Data Manipulation -SQL for Data Query -Data Updation and Deletion,

The Return of SQL, Hive, Pig- Early Database Systems-The First Database Revolution-The Second Database Revolution -The Third Database Revolution- Big Data Revolution- Big data characteristics- Introduction to Hadoop Ecosystem.

Next Generation Databases:

JSON Document Databases, XML and XML Databases, In-Memory Databases, Distributed non-relational storage systems-NoSQL Databases - Review of the Relational Model-ACID Properties-Distributed Databases: Sharding and Replication-Consistency-The CAP Theorem-NoSQL Data Models

Document Databases:

The Document Data Model-Documents and Collections-MongoDB Use Cases-Embedded Data Models-Normalized Data-Replication via Replica Sets-MongoDB Design-MongoDB and the CAP Theorem-The MongoDB Data Manipulation Language-Batch Processing and Aggregation-Indexing-MongoDB as a File System.

Column Database & Graph Database:

The Column-Family Data Model-Databases and Tables-Columns, Types, and Keys-The Data Manipulation Language-Cassandra's Architecture-Key Spaces, Replication, and Column-Families-Managing Cluster Nodes- Overview of Graph Theory- The Graph Data Model-Graph Database Use Cases-Neo4j Design: Standalone and Cluster-CRUD Operations with the Neo4j Core API-Navigating Graphs with the Traversal API-Querying as Graph Traversal.

Learning Resources

1. Sadalage, P. & Fowler, M. (2012). NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence. (1st Ed.). Upper Saddle River, NJ: Pearson Education, Inc. ISBN-13: 978-0321826626 ISBN-10: 0321826620
2. Redmond, E. & Wilson, J. (2012). Seven Databases in Seven Weeks: A Guide to Modern Databases and the NoSQL Movement (1st Ed.). Raleigh, NC: The Pragmatic Programmers, LLC. ISBN-13: 978-1934356920 ISBN-10: 1934356921
3. Dan Sullivan. NoSQL for Mere Mortals. Addison-Wesley Professional. 2015. ISBN: 0134023218 (DS) •
4. Guy Harrison. Next-Generation Databases. Apress. 2016. ISBN: 9781484213292 (GH)

SEMESTER II

DSC521 Mathematical Foundation for Data Science [2-0-0-2]

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 References:

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.

DSC523 Data Mining [3-0-0-3]

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.

Data mining algorithms: 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

Lab Experiments

- 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.

BML521 Distributed Systems for Big Data Management and Processing [1-0-0-1]

Course Objectives

- To know the fundamental concepts of big data and analytics and distributed technologies in Big Data Analytics.
- Be familiar with how to use the public domain big data tooling pipeline such as Hadoop and its eco system (Hbase, Hive, Pig, Spark)
- To introduce methods, technologies, and computing platforms for performing data analysis at distributed scale.
- To provides an in-depth understanding of terminologies and the core concepts behind big data problems.

Course Outcomes

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

- Work on map-reduce, streaming, and external memory algorithms and their implementations using Hadoop and its eco-system (Hive, Pig, Spark, Kafka).
- Students will gain practical experience in analyzing large existing databases.

Pre-Requisites

Basic programming knowledge (python), SQL and basics of statistics.

Syllabus

Big Data & Distributed Systems:

Introduction – What is Big Data?Handling and Processing Big Data- Methodological Challenges and Problems- Role of Distributed Computing in Big Data Analytics-Fundamental Concepts of Distributed Computing Used in Big Data Analytics-Distributed Computing Patterns Useful in Big Data Analytics -Distributed Technologies in Big Data Analytics- Security Issues and Challenges in Big Data Analytics in Distributed Environment.

Hadoop and Its Ecosystems:

Introduction to Hadoop-Understanding HDFS and MapReduce, Introduction - installation and execution - PIG Data Model - PIG Latin - Input, OutputRelational Operators - User Defined Functions - Join Implementations - Integrating Pig with Legacy Code and Map Reduce - Developing and Testing Pig Latin Scripts - Embedding Pig Latin in Python. Introduction to HIVE - Data Types and File Formats - Databases in Hive - HiveQL: Data Definition - Data Manipulation - Queries - Views - Indexes - Schema Design- Apache HBase.

Apache Spark:

Overview of Spark – Hadoop vs Spark – Cluster Design – Cluster Management – performance, Application Programming interface (API): Spark Context, Resilient Distributed Datasets, Creating RDD, RDD Operations, Saving RDD - Lazy Operation – Spark Jobs, Introduction of Big data Machine learning with Spark - Big Data Machine Learning Algorithms in Spark - Introduction to Spark MLlib - Introduction to Deep Learning for Big Data - Introduction to Big Data Applications (Graph Processing) -- Introduction to Graph - Introduction to Spark GraphX.

Apache Kafka:

Introduction and Features of Kafka- Kafka Usercases- Kafka Terminologies-Kafka Components-Kafka Architecture- Kafka Clusters - Kafka Producer- Kafka Consumer-Zookeeper Architecture-Kafka Installation-Implementing Single Node-Single Broker Cluster

Learning Resources

- Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale by Tom White (2015)
- Data Analytics with Hadoop: An Introduction for Data Scientists by Benjamin Bengfort and Jenny Kim (2015)
- Hadoop Application Architectures: Designing Real-World Big Data Applications by Mark Grover and Ted Malaska (2015)
- Learning Spark: Lightning-Fast Big Data Analysis by Holden Karau and Andy Konwinski (2015)
- Advanced Analytics with Spark: Patterns for Learning from Data at Scale by Sandy Ryza and Uri Laserson (2015)
- Hadoop 2 Quick-Start Guide: Learn the Essentials of Big Data Computing in the Apache Hadoop 2 Ecosystem by Douglas Eadline (2015)
- The Stratosphere platform for big data analytics by A. Alexandrov et al, The International Journal on Very Large Data Bases, pp. 939-964 (2014).
- Kafka Streams in Action: Real-time apps and microservices with the Kafka Streams API by Bill Bejeck, ISBN-13: 978-1617294471.
- Kafka: The Definitive Guide: Real-Time Data and Stream Processing at Scale by Neha Narkhede, ISBN-13: 978-1491936160.

BML522 Big Data Visualization [1-0-0-1]

Course Objectives

- To give a deep understanding of the tools and techniques for big data visualization
- To enhance the knowledge in building better visualizations using the appropriate tools and techniques
- Analyze and manipulate big data for different use cases to improve business decision making to get a competitive advantage

Course Outcomes

- Ability to manipulate and analyze big data using appropriate tools

- Ability to create visualizations that can help make better business decisions

Syllabus

Introduction: Data Visualization Process, Conventional data visualization concepts, Big data visualization – challenges, Big data categorization, visualization philosophies, Basic Charts and Plots- Multivariate Data Visualization- Data Visualization Techniques– Visualizing Complex Data and Relation, Data Visualization Tools

Hadoop and R programming: Introduction to Hadoop, Basics, log files and excel, Hadoop and Big Data, R programming, definitions and explanations, adding context, R and big data, manipulating and digging with R.

Big Data Quality: Do's and Dont's of data visualization, do not harm guide- tips on data manipulation and visualization, Data quality in big data, data manager, ensuring big data quality – examples, reformatting, consistency, reliability, appropriateness, accessibility

Big Data Tools: D3 and big data, introduction to D3, visualization using D3, Introduction to Dashboarding, different tools for dashboarding, principles and techniques used for dashboarding

Learning Resources

1. Miller, James D. Big data visualization. Packt Publishing Ltd, 2017.
2. Kirk, Andy. Data Visualization A Handbook for Data Driven Design, Sage Publications, 2016.
3. Wexler, Steve, Jeffrey Shaffer, and Andy Cotgreave. The big book of dashboards: visualizing your data using real-world business scenarios. John Wiley & Sons, 2017.
4. Wexler, Steve. The Big Picture: How to use data visualization to make better decisions - faster. McGraw Hill Education, 2021.
5. Loth, Alexander. Visual analytics with Tableau. John Wiley & Sons, 2019.
6. Härdle, Wolfgang, Henry Horng-Shing Lu, and Xiaotong Shen, eds. Handbook of big data analytics. Springer International Publishing, 2018.
7. Prajapati, Vignesh. Big data analytics with R and Hadoop. Packt Publishing Ltd, 2013.
8. Grover, Mark, et al. Hadoop application architectures: Designing real-world big data applications. " O'Reilly Media, Inc.", 2015.
9. Schwabish, Jonathan, and Alice Feng, "Do No Harm Guide: Applying Equity Awareness in Data Visualization", https://www. urban. org/research/publication/do-no-harm-guide-applying-equity-awareness-data-visualization (2021).

SEMESTER III

DSC611 Machine Learning: Principles and Practices [3-0-0-3]

Prerequisites: 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, Gausian discriminant analysis (Naïve bayes, Naïve bayes with Laplace smoothing), Binary and multiclass Classification: SVM (Quadractic 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

Experiments

- 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.

DSC612 Neural Networks and Deep Learning [3-0-0-3]

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 Knowlegde 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: Deep Feed Forward network, regularizations, training deep models, dropouts, Convolutional Neural Networks, 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, Fairness Accountability Transparency and Ethics in deep learning

Learning Resources

1. Deep Learning by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
2. Neural Networks and Deep Learning by Michael Nielson
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.

DSC613 Big Data Analytics [2-0-2-3]

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 Hodoop 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

Pre-Requisites

Basic programming knowledge (python) and basics of statistics.

Syllabus

NoSQL Database: NoSQL Databases - 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.

Lab assignments

Exploring data analysis techniques library of python

1. Experiments of various data plotting and visualization techniques
2. Programming on Hadoop
3. Programming on PIG, HIVE, and Spark
4. Implementation of Machine Learning techniques on Big Data
5. 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.

BML611 Big Data Security [1-0-0-1]

Course Objectives

- To understand the essential principles and techniques associated with Big data security.
- To understand how to examine and analyse secure data access in Big data.
- To become familiar with AI, Cryptography and Blockchain concepts in Big data security.

Course Outcomes

Students will be able to:

- Acquire Knowledge on the features and development of security methods.
- Define the principles of Privacy and Security in Big data.
- Applying AI and Cryptographic techniques to large data sets.
- Acquire Knowledge on the features of Hacking and Forensics in Bigdata.

Syllabus

Security Overview in Bigdata: Introduction to Big data, Challenges and Solutions, Security Overview, Threat categories, Hadoop Security, Security and Privacy challenges in Big data, Secure data access methods for Big data. AI and Security of critical Big data structure. Cybersecurity: Issues and Challenges in Big Data.

Big data privacy: Privacy in Big data, Issues of Privacy in Bigdata, Authentication, Authorization and accountability in Big data and Hadoop. Security in Big data with Cryptography, Block chain and Steganography: Challenges and Applications. Security Integration of Cloud with Big data,

Data protection and Big data Security Intelligence: Data protection, Data centric and ECC for Big data, Performance evolution of protocols for Big data, Intrusion and Anomaly detection in Big data, Artificial Intelligence in Big data security, Nature inspired technologies for Big data security. Existing research on AI in Big data security.

Recent trends in Big Data Security: Big data in Hacking and Social Engineering, Digital Identification protection, Big data in Surveillance security, Role of Big Data in Forensics: Opportunities and future Technologies, Case Studies.

Learning Resources

1. Fei Hu, “Big data: Storage, Sharing and Security” CRC press, Taylor and Francis Groups, 2016.
2. Indradip Banerjee, Shibakali Gupta, Siddhartha Bhattacharyya, “Big Data Security”, De Gruyter, 2019.
3. Choo, Kim-Kwang Raymond Choo, Ali Dehghantanha, “Handbook of Big Data Privacy”, Springer, 2020.
4. Ben Spivey & Joey Echeverria, “Hadoop Security: Protecting Your Big Data Platform”, O’Reilly, 2015.
5. Ramesh C. Joshi, Brij B. Gupta, “Security, Privacy, and Forensics Issues in Big Data”, IGI Global publications, 2020
6. David Lyon and David Murakami Wood, “Big Data Surveillance and Security Intelligence” UBC Press, 2021.
7. Anno Bunnik, Anthony Cawley, Michael Mulqueen and Andrej Zwitte “BIG DATA CHALLENGES: Society, Security, Innovation and Ethics”, Palgravepoint, 2016.
8. Richard Jiang Somaya Al-maadeed Ahmed Bouridane Danny Crookes Azeddine Beghdadi Editors, “Biometric Security and Privacy: Opportunities & Challenges in The Big Data Era”. Springer 2017.

SEMESTER IV

BML621 Cloud Computing for Big Data [1-0-2-2]

Course Objectives:

- To learn the various concept of Distributed and Cloud computing and to study the Architecture and service models in Cloud computing
- Optimize business decisions and create competitive advantage with Big Data.
- Derive business benefit from unstructured data

Course Outcomes:

Students will be able to:

- Acquire Knowledge on the features and development of Cloud Computing.
- Define the principles of virtualization.
- Applying data modelling techniques to large data sets
- Creating applications for Big Data analytics
- Building a complete business data analytic solution

Syllabus:

Introduction, Historical Developments, Building Cloud Computing Environments, Computing Platforms and Technologies, Virtualization: Characteristics, Environments Taxonomy, Pros and Cons, Technology Examples.

Cloud Computing Architecture: Cloud Reference Model, Types of Clouds, Economics of the Cloud, Open Challenges - Cloud Application - Cloud Programming and Management.

Introduction To Big Data - Relationship between Cloud Computing and Big Data: Overview – Benefits and Challenges – Factor facilitates big data uses in cloud - Cloud System for data management – Case study: Streaming data analysis, Tweet analysis in Social Network, Data Duplicate detection.

Big data virtualization – Resource Scheduling – Centralized to decentralized social network - Big data in Cloud Computing risks – Business challenges – Applications.

Learning Resources:

1. Michael Miller, "Cloud Computing", Dorling Kindersley India,2009.
2. Anthony T. Velte, Toby J. Velte and Robert Elsenpeter, "Cloud computing: A practical Approach", McGraw Hill,2010
3. Kai Hwang, Geoffrey C.Fox, and Jack J. Dongarra, "Distributed and Cloud Computing", Elsevier India Private Limited, 2012.
4. Rajkumar Buyya, Christian Vecchiola, and Thamarai Selvi , Mastering Cloud Computing – McGraw Hill Education.
5. Seema Acharya, Subhashini Chellappan, Big Data and Analytics –Willey India ISBN 13 9788126554782

DSC624 Natural Language Processing and Large Language Models [2-0-2-3]

COURSE PREREQUISITES

- Probability and Statistics
- Machine Learning & Deep Learning

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

LAB PROGRAMS

- 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

TEXTBOOKS/ REFERENCES

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.

BML622 Designing MLOps for enterprises [2-0-0-2]

Course Objectives

- To impart knowledge on production-level challenges of ML models
- To provide comprehension of various activities involved in the development, deployment, and monitoring of ML models
- To familiarize the principles of MLOps and different platforms

Course Outcomes

- 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

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

6. 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.
7. Burkov, Andriy. *Machine Learning Engineering.* True Positive Inc. , 2020.
8. Ameisen, Emmanuel. *Building Machine Learning Powered Applications.* O'Reilly Media, 2020.
9. Alla, Sridhar, and Suman Kalyan Adari. *Beginning MLOps with MLFlow*. Apress, 2021.
10. Rao, Dattaraj. *Keras to Kubernetes: The Journey of a Machine Learning Model to Production*. John Wiley & Sons, 2019.
11. Sculley, David, et al. "Machine learning: The high interest credit card of technical debt." (2014).
12. Jez Humble, David Farley. Continuous Delivery.

BML623 AI and ML for Big Data [1-0-0-1]

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.

BML624 Realtime Big Data Analytics [1-0-0-1]

Course Objectives

- To introduce theoretical foundations, algorithms, methodologies, and applications of streaming data.
- To familiarize students with current techniques on monitoring distributed data streams.
- To provide practical knowledge for handling and analyzing streaming data in real world problems

Course Outcomes

Upon completion of the course, the students will be able to

- Understand the applicability and utility of learning algorithms in a streaming environment.
- Describe and apply current research trends in data-stream processing.
- Analyze the suitability of stream mining algorithms for data stream systems.
- Solve problems in real-world applications that process data streams.

Syllabus

Data Streams-Characteristics - Challenges in mining data streams- Requirements and principles for real time processing- Change detection – constructing histograms from data streams

Learning from data streams - Clustering from data streams - clustering examples - clustering variables- Frequent pattern mining - frequent Itemset mining - Sequence pattern mining- Decision trees from data streams - Very Fast Decision Tree algorithm (VFDT) –Novelty detection in data streams -learning and novelty - novelty detection as a one-class classification problem

Evaluating streaming algorithms - Evaluation Issues- Evaluation Metrics- Error Estimators using a Single Algorithm and a Single Dataset- Evaluation Methodology in a Non-Stationary Environments

Tools for real time analytics- Apache Kafka- Architecture- Kafka stream processing Examples- Apache Storm- Real-time stream processing use cases

Learning Resources

1. Charu C. Aggarwal, “Data Streams: Models and Algorithms”, Kluwer Academic Publishers, Springer 2007 Edition.
2. Joao Gama, “Knowledge Discovery from Data Streams”, CRC Press, 2010.
3. Byron Ellis, “Real Time Analytics: Techniques to Analyze and Visualize Streaming Data”, John Wiley and Sons, 2014.
4. Shilpi Saxena, Saurabh Gupta, “Practical Real-time Data Processing and Analytics”, Packt publishing 2017.
5. Gwen Shapira, Todd Palino, Rajini Sivaram, Krit Petty, “Kafka – The Definitive Guide: Real-time data and stream processing at scale”, Second Edition, O’Reilly Media, 2021.
6. Ankit Jain, “Mastering Apache Storm: Processing big data streams in real time ”, Packt Publishing, 2017.

****