Anekant Education Society's

Tuljaram Chaturchand College of Arts, Science and Commerce, Baramati

(AUTONOMOUS)

Department of Statistics

Two Years Post Graduate Program

M.Sc. Data Science

Course Structure and Syllabus

Semester- I

Paper Code	Course Title	No. of Credits
PSDS111	Linear Algebra in Matlab	04
PSDS112	Probability Distributions	04
PSDS113	Optimization Techniques	04
PSDS114	Statistical Inference	04
PSDS115	Database Management System	04
PSDS116	Introduction to MATLAB and R	04

Semester- II

Paper Code	Course Title	No. of Credits
PSDS121	Design and Analysis of Experiments	04
PSDS122	Regression Analysis and Predictive Models	04
PSDS123	Statistical Quality Control	04
PSDS124	Computational Statistics	04
PSDS125	Bayesian Inference	04
PSDS126	Python and SQL Programming	04

Semester- III

Paper Code	Course Title	No. of Credits
PSDS231	Stochastic Models and Applications	04
PSDS232	Exploratory Multivariate Data Analysis	04
PSDS233	Time series analysis and Forecasting	04
PSDS234	Artificial Intelligence	04
PSDS235	Text Mining and Natural Language Processing	04
PSDS236	Data Visualization using Tableau	04

Semester- IV

Paper Code	Course Title	No. of Credits
PSDS241	Machine Learning	04
PSDS242	Discrete Data Analysis	04
PSDS243	Supply Chain & Logistics Analytics	04
PSDS244	Deep Learning	04
PSDS245	Thesis	08

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- I)
Paper Code : PSDS111
Paper : I
Title of Paper : Linear Algebra in Matlab
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

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

1. Understand basic matrix properties like rank, determinant, inverse and a special type of matrices
2. Introduce Gaussian / Gauss-Jordan elimination methods, LU factorization technique
3. Understand the concepts of vector space and subspaces.
4. Find the matrix representation of a linear transformation given bases of the relevant vector spaces.
5. Compute inner products on a real vector space and compute angle and orthogonally in inner product spaces.
6. Understand the use of linear algebra and matrices in several important, modern applications of research and industrial problems involving statistics.

Learning Outcomes:

1. Having a clear understanding of the subject related concepts and contemporary issues.
2. Having computational thinking.
3. Ability to translate vast data into abstract concepts and to understand database reasoning.
4. Having problem-solving ability- solving social issues and engineering problems.

Unit 1

Vector: Vector addition, Scalar Vector multiplication, Inner Product, Complexity of Vector Computations Vector Spaces, Subspaces, Basis and dimension of a vector space, linear dependence and linear independence, spanning set. Linear transformation, kernel, range, Matrix Representation of a linear transformation, rank nullity theorem, change of basis and similar matrices. Inner-product spaces, orthogonal sets and bases, Orthogonal Projection, Gram-Schmidt orthogonalization process. Norms and Distance, Standard deviation, Angle, Complexity. **(15 L)**

Unit 2

Algebra of Matrices, Trace and Rank of a Matrix and their properties, Determinants, Inverse, Left and right inverses, Pseudo Inverse, nilpotent matrix , idempotent matrices and their properties, Eigen values and Eigen vectors, symmetric, orthogonal , , Gauss elimination, row

canonical form, diagonal form, triangular form, Gauss-Jordan-LU decomposition, solving systems of linear equations. **(15 L)**

Unit 3

L1 norm, L2 norm, regularization of norm, covariance matrix, Singular Vectors, Singular Value Decomposition, Best Rank k Approximations, algebraic and geometric multiplicities, Cayley-Hamilton theorem, Power Method for Singular Decomposition, Singular Vectors and Eigen Vectors, Applications of Singular Value Decomposition to Centering Data ,Principal Component Analysis, Ranking Documents and Web Pages, Discrete Optimization Problem.

(15 L)

Unit 4

Generalized inverses (g-inverses), Methods of constructing g-inverses, general solution to a system of linear equations. Sparse matrices, Linear Discriminant Analysis and Canonical Correlation Analysis Spectral decomposition, Quadratic forms, definiteness and related results with proofs. **(15 L)**

References Books:

1. Bapat, R.B. (2011).Linear Algebra and Linear Models. Springer and Hindustan Book Agency.
2. Stephen Boyd (Stanford University) and Lieven Vandenberghe (University of California, Los Angeles), Introduction to Applied Linear Algebra Vectors, Matrices and Least Squares , Cambridge University Press
3. John Chandler, Swarna Reddy, Algorithms for Data Science by Brian Steele, Springer International Publishing Switzerland 2016 .Kollo, T. and Rosen, D. von (2005). Advanced Multivariate Statistics with Matrices, Springer, and New York.
4. Kumaresan, S. (2000). Linear Algebra: A Geometric Approach, Prentice Hall
5. Lay, D. C. Lay, S. R. and Mc Donald, J. J. (2016) .Linear Algebra and Its Applications, Fifth Edition, Pearson, Boston.
6. Ramachandra Rao, A. and Bhimasankaram, P. (2000). Linear Algebra. Hindusta Book Agency
7. Rao, C. R. (1995). Linear Statistical Inference and Its Applications, Wiley
8. Searle, S. R. (1982). Matrix Algebra Useful for Statistics, John Wiley, New York.
9. G. Allaire and S. M. Kaber. Numerical Linear Algebra, Texts in Applied Mathematics, Springer, 2008. L. Hogben, Handbook of Linear Algebra, CRC Press/Taylor & Francis Group, 2014. Friedberg, S., Insel, A., and Spence, L., Linear Algebra, 5/e, Pearson, 2019.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- I)
Paper Code : PSDS112
Paper : II
Title of Paper : Probability Distributions
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To incorporate the concepts of probability theory and its applications as the core material in building theoretical ideas along with the practical notion.
2. To integrate the intrinsic ideas of preliminary and advanced distributions to correlate with the real-world scenarios.
3. To understand various types of generating functions used in statistics.

Learning Outcomes:

Student will be able to:

1. Develop problem-solving techniques needed to calculate probability and conditional probability.
2. Formulate fundamental probability distribution and density functions, as well as functions of random variables, derive the probability density function of transformations.
3. Derive the expectation and conditional expectation, and describe their properties.

TOPICS/CONTENTS:

Unit 1

Probability and Random variables

Introduction – Random Experiments, Empirical basis of probability, Algebra of events, laws of Probability; Conditional Probability, Independence, Bayes' law; Application of probability to business and economics. One-dimensional Random variable- Discrete and Continuous; Distribution functions and its properties; Bivariate Random Variables- Joint Probability functions, marginal distributions, conditional distribution functions, Notion of Independence of Random variables **(15 L)**

Unit 2

Functions of Random Variables

Functions of random variables: introduction, distribution function technique, transformation technique: one variable, transformation technique: several variables, theory and applications.

Mathematical Expectation

Expectation, Variance, and Co-variance of random variables; Conditional expectation and conditional variance; Markov, Holder, Jensen and Chebyshev's Inequality, Weak Law of Large numbers, Strong law of large numbers and Kolmogorov theorem, Central Limit Theorem. **(15L)**

Unit 3

Generating Functions

Probability generating function (p.g.f.), moment generating function (m.g.f.), characteristic function (c.f.) Properties and Applications. Probability distributions of functions of random variables: one and two dimensions.

Sampling Distributions

Introduction, The sampling distribution of the Mean: Finite Populations, Sampling distribution of the proportion: Finite Populations, distribution of sample variance, the chi-square distribution, the t distribution, the F distribution, order statistics: properties, and applications, procedure of hypothesis testing. **(15L)**

Unit 4:

Discrete Distributions

Bernoulli, Binomial, Poisson, Geometric, Hypergeometric, Negative Binomial, Multinomial, distributions and Discrete Uniform distribution - definition, properties and applications with numerical problems.

Continuous Distributions

Uniform, Normal distribution function, Exponential, Gamma, Beta distributions (First and Second kind), Weibull, Cauchy and Laplace distributions, lognormal, logistic, Pareto, Chi-sqaure and Rayleigh distribution functions - definition, properties and applications; concept of truncated distributions. **(15L)**

References Books:

1. Parimal Mukhopadhyay; An Introduction to the Theory of Probability, World scientific, 2012.
2. Irwin Miller, Marylees Miller, John E. Freund's; Mathematical Statistics, Pearson, 2017.
3. Fetsje Bijma, Marianne Jonker and Aadvander Vaart; Introduction to Mathematical Statistics, Amsterdam University Press, 2018.
4. Krishnamoorthy, K., Handbook of Statistical Distributions with Applications, Chapman & Hall/CRC, 2006.
5. Rohatgi, V.K. and Ebsanes Saleh, A. K. Md., An introduction to Probability and Statistics, 2nd Ed., John Wiley & Sons, 2002.

6. Shanmugam, R., Chattamvelli, R. Statistics for scientists and engineers, John Wiley, 2015.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- I)
Paper Code : PSDS113
Paper : III
Title of Paper : Optimization Techniques
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To familiarize the students with some basic concepts of optimization techniques and approaches.
2. To formulate a real-world problem as a mathematical programming model.
3. To develop the model formulation and applications are used in solving decision problems.
4. To solve specialized linear programming problems like the transportation and assignment Problems.

Learning Outcomes:

Students are expected to

1. Students can identify and develop operational research models from the verbal description of the real system.
2. Students will be able to understand the characteristics of different types of decision-making environments and decision-making approaches.
3. Students can apply optimization techniques to take correct decision.

TOPICS/CONTENTS:

Unit 1
Introduction to Operations Research

Introduction to Mathematical models of Operation Research, Scope and applications of Operation Research, Phases of Operation Research, Characteristics of Operation Research, and Limitations of Operation Research**.**

Linear Programming

Introduction, Properties of Linear Programming, Basic assumptions, Mathematical formulation of Linear Programming, Limitations or constraints, Methods for the solution of LP Problem, Graphical Method, Simplex Method, Big M Method. **(15)**

Unit 2
Dual Linear Programming

Introduction to Primal and Dual problem, Dual problem properties, Solution techniques of Dual problem, Dual Simplex method, Relations between direct and dual problem, Economic interpretation of Duality. **(15L)**

Unit 3

Transportation and Assignment Models

Introduction to transportation problem (TP), Balanced and Unbalanced TP, Methods of basic feasible solution, Optimal solution, MODI method. Assignment problem, Hungarian Method.

Network Analysis

Basic concepts, Construction of Network, Rules and precautions, CPM and PERT Networks, obtaining of critical path, probability and cost consideration, advantages of Network. **(15L)**

Unit 4

Theory of Games

Introduction and terminology of Two Person Zero-Sum Game, Solution of games with saddle points and without saddle points, 2×2 games, dominance principle, m×2 and 2×n games, Graphical method.

Hyper parameter optimization

Gradient of a function, Steepest descent method, Nelder Mead's Simplex search method, Newtons method. **(15L)**

References Books:

1. S.D. Sharma (2000), Operations Research, Nath& Co., Meerut.MauriceSolient, Arthur Yaspen, Lawrence Fridman, (2003), OR methods and Problems,New Age International Edition.
2. J K Sharma (2007), Operations Research Theory & Applications, 3e, Macmillan India Ltd. P. SankaraIyer, (2008), Operations Research, Tata McGraw-Hill.
3. Taha, H.A., Operations Research: An Introduction, Prentice Hall of India, 9th Edition,2010.
4. A Ravindran, Don T Philips and James J Solberg, Operations Research: Principles and Practice, 2nd edition, John Wiley and sons, 2007.
5. L.S.Srinath, PERT and CPM Principles and Applications, Affiliated East-West Press(Pvt.) Ltd, 3rd edition, 2001

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- I)
Paper Code : PSDS114
Paper : IV
Title of Paper : Statistical Inference
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

Students successfully completing this course will be able to:

1. Get basic understanding about statistical models and their use.
2. Apply linear and regression models depending upon the problem context.
3. Get a better understanding of probabilistic models.
4. Implement different tree based models.
5. Derive inference from different statistical data sets.

Learning Outcomes:

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

1. Learn the approaches to point estimation of parameters.
2. Understand the concept of interval estimation and confidence intervals.
3. Basic concepts in tests of hypotheses.

Unit 1

Introduction

Population, sample, parameter and statistic; characteristics of a good estimator, Unbiasedness, Sufficiency – Factorization Theorem – Minimal sufficiency, Efficiency – Most Efficient estimator, likelihood equivalence, applications of Lehmann-Scheffe's Theorem, Rao-Blackwell Theorem and, Consistency –Invariance property of Consistent estimator, sufficient condition for consistency uniformly minimum variance unbiased Estimator. **(15 L)**

Unit 2

Point Estimation

Point Estimation- Estimator, Estimate, Methods of point estimation – Maximum likelihood Method (the asymptotic properties of ML estimators are not included), and large sample Properties of ML estimator (without proof) -applications, Method of moments, method of Least squares, method of minimum chi-square and modified minimum chi-square asymptotic, Maximum Likelihood Estimation and applications.

(15 L)

Unit 3

Testing of hypotheses

Types of errors, power of a test, most powerful tests, Neyman–Pearson Fundamental Lemma and its applications; Notion of Uniformly most powerful tests; Likelihood Ratio tests, Description and property of LR tests - Application to standard distributions.

Large sample tests

Large sample properties, Tests of significance (under normality assumption) Test for a Population mean, proportion, Test for equality of two means, proportions, Test for Variance, Test for correlation, Test for Regression.

Small sample tests

Student's t-test, test for a population mean, equality of two population means, paired t-test, F-test For equality of two population variances, Chi-square test for goodness of Fit and test for independence of attributes, χ^2 test for testing variance of a Normal Distribution Analysis of Variance.

Unit 4

Interval estimation confidence limits and confidence coefficient;

Duality between acceptance region of a test and a confidence interval, Construction of confidence intervals for population proportion (Small and large samples) and between two population proportions (large samples), Confidence intervals for mean and variance of a normal population; Difference between the Mean and ratio of two normal populations.

Non-parametric tests Sign test, Signed rank test, Median test, Mann-Whitney test, Run test and one sample Kolmogorov – Smirnov test, Kruskal – Wallis H test (Description, properties and applications only). **(15 L)**

All topics to be covered using R software. Manual calculations are not expected.

Reference Books:

1. Casella G. and Beregar R.L. (2002) Statistical Inference, 2nd Edition (Duxbury Advanced Series)
2. Dudewitz E.J. & Mishra S.N.(1988) Modern Mathematical Statistics (John Wiley)
3. Kale B.K. (1999) A First course on Parametric Inference (Narosa)
4. Lehman E.L (1988) Theory of point estimation (John Wiley)
5. Lehman E.L(1986) Testing of Statistical hypotheses (John Wiley)
6. Rohatagi V.K. (1976) Introduction to theory of probability & mathematical statistics (John Wiley & sons)

7. Dasgupta A. (2008), Asymptotic Theory of Statistics and Probability, Spinger-Verlag, New York.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class	: M. Sc. (Semester- I)
Paper Code	: PSDS115
Paper	: V
Title of Paper	: Database Management System
Credit	: 4 credits
No. of lectures	: 60

Learning Objectives:

Students successfully completing this course will be able to:

1. Recognize database system design and implementation.
2. Examine the logical, physical, and database modelling designs..
3. Developing an understanding of key DBMS principles.
4. Recognize how to create, modify, and query databases for data.

Learning Outcomes:

1. Be familiar with the fundamentals of database concepts and database management systems.
2. Utilize conceptual modelling techniques, like as the ER model and relational model, to model the data requirements for an application.
3. Write SQL commands to create tables, insert, update, delete and querying data.

Unit 1

Introduction to File organization &DBMS, Database-system Applications, Purpose of Database Systems, Types of file Organization, File system Vs. DBMS, Data models, Levels of abstraction, Data in dependence, Structure of DBMS, Users of DBMS, Database Architecture, Speciality Databases.

(15L)

Unit 2

Structure of Relational Databases, Database Schema, Keys, Relational Operations, Conceptual Design (E-R model), Overview of DB design, ER data model (entities, attributes, entity sets, relations, relationship sets),Additional constraints (Key constraints, Mapping constraints), Conceptual design using ER modelling. Relational data model, Conversion of ER to Relational model, Integrity constraints, Relational algebra, Preliminaries. **(15L)**

Unit 3

Introduction to SQL, Basic structure, Set operations, Aggregate functions, Null values , PL/PgSqL: Data types, Language structure, Operations with SQL, Nested Sub queries, Modifications to Database, DDL and DML commands with examples, SQL mechanisms for joining. **(15L)**

Unit 4

Intermediate and advanced SQL: Join Expressions- Join conditions, Outer joins, Join types and conditions, Views- View definition, using views in SQL queries, Materialized views, update a view 4.3 Create table extensions, Schemas, Catalogs and Environments, The relational Algebra, The tuple relational calculus. **(15L)**

Reference Books:

1. Abraham Silberschatz, Henry F. Korth, S. Sudarashan, Database System Concepts, McGraw-Hill International Edition, Sixth Edition
2. Elmasri, Navathe, Fundamentals of Database Systems, Pearson Education, Third Education
3. Ramakrishnan, Gehrke, Database Management Systems, McGraw Hill International Edition, Third Edition
4. Peter Rob, Carlos Coronel, Database System Concepts, Cengage Learning, India Edition
5. S.K. Singh, "Database Systems Concepts, Design and Applications", First Edition, Pearson Education, 2006
6. Redmond, E. & Wilson, Seven Databases in Seven Weeks: A Guide to Modern Databases and the No SQL Movement Edition:1st Edition.
7. Shamkant B. Navathe, Ramez Elmasri,(2010), Database Systems, ISBN:9780132144988, PEARSON HIGHER EDUCATION
8. Richard Stones, Neil Matthew, (2005), Beginning Databases with PostgreSQL: From Novice to Professional, ISBN:9781590594780, Apress
9. Korry, Douglas, (2005), Postgre SQL, ISBN:9780672327568, Sams Publishing.
10. Joshua D. Drake, John C. Worsley, Practical Postgre SQL, (2002), ISBN:9788173663925 O'Reilly Media, Inc., ISBN: 9781565928466

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- I)
Paper Code : PSDS116
Paper : VI
Title of Paper : Introduction to MATLAB and R
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To gain knowledge about MATLAB and R software for linear algebra, probability distribution and statistical inference.

Learning Outcomes:

1. Students will be able to solve linear algebra problems using MATLAB/R - software.
2. Students will be able to draw model sample from distributions.
3. Students will understand asymptotic behaviour of the estimators, find and verify the consistent estimator and consistency.

Sr. no.	Title of Experiments
1.	Calculation of determinant and rank of higher order by partitioning method
2.	Calculation of inverse of matrices of higher order.
3.	Calculation of Moore-Penrose inverse and g-inverse.
4.	Solution of simultaneous equations.
5.	Eigen value, Eigen vectors, Spectral decomposition, Power of matrix
6.	Plotting likelihood function and obtaining MLE.
7.	Plot density function, Distribution function and computation of probability of events related to standard probability distribution.
8.	Model sampling from Gamma, Chi square, Weibull, Lognormal probability distribution
9.	Model sampling from Discrete and continuous distribution.
10.	Verification of Consistency and CAN estimator.
11.	Comparison of consistent estimator based on MSE and sample size.
12.	Power function of large sample test. (LR, Wald, Rao)

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- II)
Paper Code : PSDS121
Paper : I
Title of Paper : Design and Analysis of Experiment
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. The main objective of this course is to learn and understand various designs of experiments.
2. Students should be able to design and carryout various experiments and analyse the data.
3. Students should be able to apply appropriate design in real life situation.

Learning outcomes:

1. Students will be able to understand basic principles and various terms of Design of Experiments.
2. Students will be able to apply Factorial design, fractional factorial design, confounding in real life problems.
3. Students should be able to analyse the data of various experimental design.

TOPICS/CONTENTS:

Unit 1

Analysis of one-way classification modal. Analysis of two-way classification model with equal number of observations per cell with and without interactions. Analysis of two-way classification model with unequal number of observations per cell without interactions. Introduction of the Design of Experiments (DOE), The Basic Principles of DOE, Steps for Planning, Conducting and Analysing an Experiment, Principles of scientific experimentation – Basic Designs: Completely Randomized Design (CRD), Randomized Block Design (RBD) and Latin Square Design (LSD) – Analysis of RBD (with one observation per cell, more than one but equal number of observations per cell). **(15 L)**

Unit 2

Multiple Comparisons, Multiple Range Tests, Statistical analysis of Covariance. Analysis of non- normal data using: square root transformation for counts, $Sin^{-1}(.)$ transformation for proportions, Kruskal Wallis test. **(5 L)**

Unit 3

Full Factorial experiments and their analysis, concepts of main effects, interaction effect, their graphical representation, analysis of single replicate and more than one replicates of

2^kdesignand partial confounding of 2^k,fractional experiments in 2^k. Statistical analysis of single replicate and more than one replicates of 3^kdesign, confounding and fractional experiments in 3^k. **(15 L)**

Unit 4

Balanced Incomplete Block Design (BIBD) – Types of BIBD – Simple construction methods Concept of connectedness and balancing – Intra Block analysis of BIBD. Partially Balanced Incomplete Block Design with two associate classes – intra block analysis. Split plot and strip plot design and their analysis. **(15 L)**

Unit 5

Response surface methodology (RSM): linear and quadratic model, stationary point, central composite designs(CCD), ridge systems, multiple responses, concept of rotatable designs, Box-Behnken design, optimality of designs, simplex lattice designs, simplex centroid designs. **(10 L)**

References Books:

1. Dean, A. and Voss, D. (1999). Design and Analysis of Experiments, Springer.
2. George E. P. Box, Draper N.R. (1987). Empirical Model-Building and Response Surfaces, Wiley.
3. Kshirsagar A.M. (1983). Linear Models, Marcel Dekker.
4. Montgomery, D.C. (2001). Design and Analysis of Experiments, Wiley.
5. Phadke, M.S. (1989). Quality Engineering using Robust Design, Prentice Hall,Englewood Cliffs, New Jersey.
6. Wu, C.F. Jeff and Hamada M. (2000). Experiments: Planning, Analysis and Parameter Design Optimization, John Wiley and Sons.
7. Bapat, R. B. (2012). Linear algebra and linear models. Springer Science & Business.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- II)
Paper Code : PSDS122
Paper : II
Title of Paper : Regression Analysis and Predictive Modelling
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. Gain knowledge of model construction and regression analysis.
2. Give the capability of developing relationships between variables
3. Investigate possible diagnostics in regression techniques

Learning Outcome:

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

1. Deep comprehension of the linear and nonlinear regression models.
2. Demonstrate understanding of model selection and regression modeling approaches.
3. The connections between dependent and independent variables should be examined.
4. Estimate the parameters and fit a model.
5. Investigate possible diagnostics in regression modeling and analysis.
6. Validate the model using hypothesis testing and confidence interval approach.

Unit 1

Simple Linear Regression Analysis:

Simple linear regression model, Ordinary Least Square method, generalized and weighted least squares, validating simple regression model using t, F test, developing confidence interval.

(10 L)

Unit 2

Multiple linear Regression Analysis

Concept of Multiple regression model, Ordinary Least Square method, generalized and weighted least squares, Assessing the fit of the regression line, inferences from multiple regression analysis, problem of over fitting of a model, comparing two regression model, prediction with multiple regression equation.

(15 L)

Unit 3

Model Adequacy Checking and Transformation Techniques:

Residual analysis, PRESS statistics, detection and treatment of outliers, lack of fit of the regression model, test of lack of fit, Problem of autocorrelation and heteroscedasticity. Variance stabilizing transformations, transformations to linearize the model, Box-Cox methods, transformations on the repressors variables. Multicolinearity, sources of multicollinearity, effects of multicollinearity. Multicollinearity diagnostics: examination of correlation matrix, variance Inflation factors (VIF), Eigen system analysis of X'X. Methods of dealing with Multicollinearity: **(18 L)**

Unit 4

Polynomial regression, Non-linear regression: Non-linear least squares transformation to a linear model, their uses and limitations, examination of non-linearity, initial estimates, iterative procedure, Newton-Raphson method. Generalized linear model: Link function: normal, binomial, Poisson, exponential, gamma. Logistic regression: Logit transform, ML estimation, tests of hypothesis, Wald test, LR test, score test, test for overall regression.

(17 L)

Reference Book:

1. Draper, N. R. and Smith H. (1998) Applied regression analysis 3rd edition (John Wiley)
2. Hosmer, D. W. and Lemeshow, S. (1989) Applied logistic regression (John Wiley)
3. McCullagh, P. and Nelder, J. A.(1989) Generalized linear models (Chapman and Hall)
4. Montogomery D. C., Elizabeth a. Peck, G. Geoffrey.(2003) Introduction to linear regression analysis (Wiley Eastern)
5. Neter, J.; Wasserman, W. and Kutner, M.H.(1985) Applied linear statistical models
6. Ratkowsky, D. A.(1983) Nonlinear regression modeling (Marcel Dekker)

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- II)
Paper Code : PSDS123
Paper : III
Title of Paper : Statistical Quality Control
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. Students should be able Understand different control chart.
2. Students should be able use the methods of statistical process control.
3. To link and analyze the various sampling schemes to find the plan for quality inspection.

Learning Outcomes:

Students will be able to

1. describe the DMAIC processes.
2. perform analysis of process capability and measurement system capability.
3. demonstrate the ability to design, use, and interpret multivariate control chart, synthetic control chart, non-parametric control chart.

TOPICS/CONTENTS:

Unit 1

Six sigma

Concept of six sigma, methods of six sigma, DMAIC methodology, DFSS methodology, six-sigma control chart, case studies.

Control Chart:

Revision of control charts for attributes, S^2chart, X -S chart with subgroup size (i) fixed, (ii) variable. Equivalence between control chart and testing of hypothesis problem. Operating characteristic (OC curve) of control chart. Average run length (ARL).Probability of false alarm, probability of catching shift in parameter. Comparison of control chart using ARL and OC curve. Patterns on control charts with justification and its effect on probability of false alarm. An application of control charts situations other than manufacturing.

Attribute control charts:

Revision of control charts for attributes, OC curve for P chart and C chart. Determination sample size for P chart by various criteria (i) probability of catching at least 0.5 (ii) to get LCL >0 (iii) To have at least some defectives in sample with given confidence coefficient. (iv) Minimizing ATS () chart and OC Curve, U chart, Demerit control chart for number of

defects. Nelsons control chart for low defect counts. General ideas of economic designing of control charts. Duncan's model for the economic control chart. **(15L)**

Unit 2

Process Capability analysis:

Meaning, Estimation technique for capability of a process –Capability Indices: Process capability ratios Cp, Cpk, Cpm, Cmk, Cpc – Process capability analysis using a control chart – Process capability analysis using design of experiments

CUSUM chart:

Chart statistic (Ci+, Ci-)and chart parameters(k, h), construction and working of tabular CUSUM chart for mean and variance, Statement of hypotheses. Estimation of shift in mean of process, fast initial response or head start feature, Sigmund's approximation for ARL and determination of chart parameters. CUSUM chart for subgroup size n>1, comparison between Shewhart chart and CUSUM chart V mask procedure.

EWMA chart:

Chart statistic its expectation and variance. Choice of chart parameters (L). Construction and working of EWMA chart for mean and variance. EWMA chart for subgroup size > 1, Comparison of Shewhart control charts with CUSUM charts. Simulation of ARL. **(15L)**

Unit 3

Acceptance sampling –

Terminologies – Attribute sampling plan by attributes – Single sampling plan and Double sampling plan – OC, ASN, AOQ, AOQL and ATI curves –MILSTD -105E Tables

Double specification limits –

M-method, Double sampling by variables - MILSTD -414 Tables – Continuous Sampling plan – CSP-1, CSP-2, CSP-3, Wald and Wolfowitz SP-A.

Attribute Sampling plans

Producers risk, Consumers Risk, designing single sampling plan for stipulated Producers and consumers risk, OC curves under Normal, Tightened and reduces inspection, Single, Double and Multiple sampling plans in AQL systems **(15L)**

Unit 4

Other control charts

i. Synthetic control chart: Concept of run length, probability distribution of run length Confirming run length (CRL) chart for attributes, Synthetic control chart, computations of chart parameters for given ARL (0), Zero State Performance, Steady state performance, Computations of ARL (), ATS ()., Comparison of with Shewhart control chart and CUSUM charts.

ii. Non-parametric control chart: Concept, construction of non-parametric chart usingsign test.

ii. A distribution-free Shewhart Quality Control Chart Based on Singed-Rank

iii. Control charts for auto correlated observations: Need, constructions of control chart for residuals after fitting first order auto correlated model.

Hotelling T^2Chart:

Testing multivariate normality, Hotelling T^2multivariate control chart for mean vector when (i) dispersion matrix is (i) known (ii) unknown ARL (0), ARL (δ). Control chart for dispersion matrix when mean vector is (i) known (ii) unknown. T^2control chart when subgroup size n=1 **(15L)**

References Books:

1. Edward G. Schilling, Dean V. Neubauer, Acceptance Sampling in Quality Control, Second Edition, Taylor & Francis, 2009
2. Poornima M. Charantimath, Total Quality Management, 3/E, Pearson India Limited, 2017.
3. Eugene L. Grant Richard S. Leavenworth, Statistical Quality Control,7 edition, McGraw Hill Education, India, 2017.
4. Douglas C. Montgomery, Introduction to Statistical Quality Control, Seventh Edition, John Wiley and Sons, New York. 2013.
5. Wu, Yeu and Spedding (2001) A synthetic control chart for detecting fraction non confirming increases JQT Vol. 33 (1), 104-111

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- II)
Paper Code : PSDS124
Paper : IV
Title of Paper : Computational Statistics
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To familiarize students to learn the design of the algorithm used to apply statistical techniques like bootstrapping and simulation.
2. To identify problems that computational statistics can be used to solve include those involving optimization and resampling techniques.
3. To acquaint students with numerical integration and the simulation of random processes or variables.

Learning Outcomes:

1. Students can employ computational techniques to provide numerical solutions to statistical questions that are challenging or unsolvable analytically.
2. Students can apply numerical techniques for transformations, for function approximation.
3. Students will be able to understand and implement the Monte Carlo Studies in Statistics and random number generators.

TOPICS/CONTENTS:

Unit 1

Theory of inverse transformation method (ITM) for random variate generation- definition of quantile function, its properties. Quantile function as a random variable and its distribution function. ITM based algorithms to generate random variates from standard discrete and continuous distributions. Generation of random variates using the relationships between distributions, composition and convolution methods. Algorithms for random variate generation from mixture distributions, Chi-square, t and F-distributions. Random variate generation from bivariate and conditional distributions. Theory of random number generation, testing a random number generator- run test, Kolmogorov-Smirnov test, sign test, rank test. Selection of a random number generator. **(15 L)**

Unit 2

Solutions to Non linear equations: Bisection method, Newton Raphson, Steepest descent, Quadrature interpolation, Jacobi and Gauss Seidel Methods. Simple Optimization method. Direct search, grid search, Hooke & Jeeves method Interpolatory search, Gradient search.

(15 L)

Unit 3

Numerical Differentiation: Forward and backward Difference, Error analysis: True solution, Approximate numerical solution, Causes of error. Numerical Integration: Trapezoidal rule, Simpson's Rule. Jack-knife and Bootstrap sampling. Bias and standard errors, Bootstrapping for estimation of sampling distribution. Confidence intervals, variance stabilizing transformation.

(15 L)

Unit 4

Methods to compute integrals- quadrature formula, double integration, Gaussian integration, Monte Carlo Methods: Monte Carlo integration and its application to compute expected values and probabilities, Verification of WLLN, CLT and other approximations through simulation.

(15 L)

References Books:

1. Atkinson K. E. (1989): An Introduction to Numerical Analysis. (Wiley)
2. Devroye L. (1986) : Non- Uniform Random Variate Generation. (Springer- Verlag New York)
3. Ephron B. and Tibshirani. R. J. (1994): An Introduction to the Bootstrap. (Chapman and Hall) 4. Morgan B. J. T.(1984) : Elements of Simulation. (Chapman and Hall)
4. Robert C. P. and Casella G. (1999): Monte Carlo Statistical Methods. (Springer Verlag New York, Inc.)
5. Ross. S. M. (2006): Simulation. (Academic Press Inc)
6. Rubinstein, R. Y. (1998) Modern Simulation and Modelling. (Wiley Series in Probability and Statistics)
7. William J., Kennedy, James E. Gentle. (1980): Statistical Computing. (Marcel Dekker)

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- II)
Paper Code : PSDS125
Paper : V
Title of Paper : Bayesian Inference
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. Introduction to Bayesian inference, Bayesian approach for data analysis in a variety of applications.
2. To obtain posterior distributions for the proportion and mean.
3. To Construct Bayesian prediction intervals and write appropriate conclusions.

Learning Outcomes:

1. Students can To Construct Bayesian prediction intervals and write appropriate conclusions.

TOPICS/CONTENTS:

Unit 1

Subjective and frequentist probability, Bayesian inference set up, prior and posterior distributions, loss functions, principles of minimum expected posterior loss, quadratic and other loss functions, advantages of being Bayesian, improper priors, Common problems of Bayesian Inference, point estimation, HPD confidence intervals, predictions of future observations, Bayesian testing. **(18L)**

Unit 2

Bayesian analysis with subjective priors, classes priors, conjugate class of priors, Jeffrey's prior, probability matching prior, robustness and sensitivity. **(12L)**

Unit 3

Bayesian model selection BIC, Bayes factors, limit of posterior distributions, consistency and asymptotic normality of posterior distributions. **(12L)**

Unit 4

Bayesian computing, E-M Algorithm, MCMC, MH Algorithms, Gibb' sampling, convergence diagnostics. (Note: Minimum 10 hours of computational practice) **(18L)**

Reference Books:

1. Bayesian Data Analysis, by Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. CRC Press/Taylor & Francis, 2013, 3rd Edition. ISBN: 9781439840955
2. Bayesian Computation with R, by Jim Albert. Springer, 2009, 2nd Edition. ISBN: 0387922970
3. A First Course in Bayesian Statistical Methods, Peter D. Hoff, 2009, New York: Springer
4. Bayesian Data Analysis. Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., & Rubin, D.B. (2013). CRC press.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- II)
Paper Code : PSDS126
Paper : VI
Title of Paper : Python and SQL Programming
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To understand why Python is a useful scripting language for developers.
2. To learn how to design and program Python applications.
3. To learn how to write loops and decision statements in Python.
4. To learn how to use class inheritance in Python for reusability.
5. To learn how to use exception handling in Python applications for error handling.

Learning Outcome:

When students complete Intro to Programming with Python, they will be able to:

1. Build basic programs using fundamental programming constructs like variables, conditional logic, looping, and functions.
2. Work with user input to create fun and interactive programs.
3. Learn about SQL – Structured Query Language ϖ Build database using Data Definition Language Statements Perform basic CRUD operations using Data Manipulation Language statements like Insert, Update and Delete Write and call Stored Procedures and Functions stored in database. Write and manage database triggers, cursors and Index.

Sr. No.	Title of Experiments
1.	Basics of Python Language When and why to use Python for Analytics • Introduction & Installation of Python • Python Syntax, Strings, Lists and Dictionaries • Loops • Regular Expressions
2.	Scientific Libraries in Python • Numpy, Scipy
3.	Introduction to Pandas • Selecting data from Pandas Data Frame, Slicing and dicing using Pandas • GroupBY / Aggregate, Strings with Pandas, Cleaning up messy data with Pandas, Dropping Entries, Selecting Entries

4.	Data Manipulation using Pandas • Data Alignment • Sorting and Ranking • Summary Statistics • Missing values • Merging data • Concatenation • Combining Data Frames • Pivot • Duplicates • Binning
5.	Data visualization on using matplotlib and seaborn libraries • Scatter plot, Line plot, Bar plot, Histogram, Box plot, Pair plot
6.	Control structures using Toyota Corolla dataset • if-else family, for loop, for loop with if break, while loop
7.	Introduction to Database Management System This module introduces you to the database, the need for databases, and their examples. Further, you will learn about Database Management Systems and its history. Lastly, you will go through various Database Management System softwares
8.	Types of Database Management System • This chapter will cover various types of DBMS, including Hierarchical, Network, Relational, and Object-Oriented Databases. You will also be familiarized with several advantages of DBMS.
9.	Introduction to SQL • This chapter will brief you on the introduction to SQL and how to install it on your system. • In this chapter, you will learn how to implement various types of Commands in MySQL, such as DDL, DQL, DML, DCL, and TCL, with hands-on demos.
10.	Filter Record in MySQL • In this chapter, you will learn how to filter the records using the WHERE clause in MySQL, Operation in MySQL
11.	Pattern Matching in MySQL Here, you will learn to find patterns using the LIKE operator with the WHERE clause
12.	Null Values in MySQL In the final chapter, you will learn to insert a new record or update a record to an optional field without adding a value with the help of a Null value

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- III)
Paper Code : PSDS231
Paper : I
Title of Paper : Stochastic Models and Applications
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To understand discrete and continuous Markov chains models to compute the probability of events.
2. Formulate and solve problems by computing the long-term probabilities of a Markov chain model.
3. Write Python/R code to simulate Markov chains, and compute probabilities of events that may be difficult to derive by hand.
4. Apply Poisson processes to model the occurrence of events in various applications.

Learning Outcomes:

1. Students will be able to formulate tpm, n-step transition probabilities
2. Students will be able to classify of states.
3. Students will perform stochastic simulations.
4. Students will become familiar with stochastic processes, including Poisson process, Wiener process and Renewal process, etc.

TOPICS/CONTENTS:

Unit 1

Notion of stochastic processes, Markov chain, one step transition probabilities, Chapman-Kolmogorov equations, evaluation of higher step transition probabilities, classification of states, periodicity of a Markov chain, concept of closed class, minimal closed class, stationary distribution. Some examples such as gamblers ruin problem and one-dimensional random walk. Concept of absorption probabilities, Use of these to compute probability of winning the game by a gambler having initial capital ‘a’

Unit 2

Branching process, classification of states, identification of criticality parameter, extinction probability, relationship between criticality parameter and extinction probability of the process, Expression for mean and variance of the process. Extinction probability, Some epidemiological applications, Introduction to Markov chain in continuous time, concept of

intensity rate, relationship between intensity matrix and transition probability matrix. Kolmogorov's forward and backward equations

Unit 3

Introduction to birth process, birth and death process, linear birth and death process, Growth model with immigration and related results, Expression for mean and variance of a birth process and, birth and death process, Applications of these processes.

Unit 4

Poisson process, two definitions and their equivalence, Distribution of inter arrival times, conditional joint distribution of inter arrival times. Compound Poisson process, Some applications. Introduction to renewal process, relationship with Poisson process, key and elementary renewal theorems associated with renewal processes.

Books Recommended

1. Bhat B.R. (2000). Stochastic Models: Analysis and Applications, New Age International.
2. Medhi, J. (2010) Stochastic Processes, New Age Science Ltd.
3. Pinsky M. A. and Karlin, S. (2010). An Introduction to Stochastic Modeling, 4thEdn. Academic Press.
4. Ross, S. (2014). Introduction to Probability Models, 11th Edn. Academic Press.
5. Feller, W. (1972). An Introduction to Probability Theory and its Applications, Vol. 1, Wiley Eastern.
6. Hoel, P.G. Port, S.C. & Stone, C.J. (1972). Introduction to Stochastic Processes, Houghton Mifflin
7. Karlin, S & Taylor, H.M. (1975). A First Course in Stochastic Processes (Second. Edition), Academic Press.
8. Serfozo, R. (2009). Basics of Applied Stochastic Processes, Springer.

SYLLABUS (CBCS) FOR M.Sc. Data Science
(With effect from Academic Year 2022-2023)

Class : M. Sc. Data Science (Semester- III)
Paper Code : PSDS232
Paper : II
Title of Paper : Exploratory Multivariate Data Analysis
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To develop feasible solution of real-life problems, using multivariate methods and techniques.
2. To develop an understanding of appropriate and relevant methods of multivariate data analysis.
3. To summaries and synthesize datasets using simple graphs, Make graphical displays of very high dimensional data
4. To use visualization methods adapted to multidimensional exploratory analysis.
5. To recognize the method adapted to the exploration of a dataset according to the nature and structure of the variables.

Learning Outcomes:

A student is able to perform analysis of multivariate data.

TOPICS/CONTENTS:

Unit 1

Multivariate data and their diagrammatic representation, Exploratory multivariate data analysis, sample mean vector, sample dispersion matrix, sample correlation matrix, graphical representation, mean, variance, co-variance, correlation of Linear transformations., six step approach to multivariate model building. Introduction to multivariate regressions models Principal component Analysis (by using covariance and correlation method, standardized method), Factor analysis (models, rotation types), Canonical correlation with real life examples. **[15L]**

Unit 2

Cluster analysis (Hierarchical and Non hierarchical, Agglomerative, Single, complete, average, Wald's linkage, K mean clustering method, qualitative method clustering).Multivariate normal distribution, Singular and non singular normal distribution, mean, variance of multivariate normal distribution, Random sampling from multivariate normal distributions, independence of variables, M.G.F. Characteristic function, moments,

Distribution of linear and quadratic form of normal variables, marginal and conditional distribution, multiple and partial correlation coefficient (3 random variable case) with examples on each of the topic. **[15L]**

Unit 3

Multivariate Linear Model and Analysis of Variance and Covariance: Maximum likelihood estimation of parameters, tests of linear hypothesis, distribution of partial and multiple correlation coefficients and regression coefficients. Multivariate linear regression, multivariate analysis of variance of one and two way classification data (only LR test). Multivariate analysis of covariance. Hotelling $T^2$and Mahalanobis D^2applications in testing and confidence set construction.

[10L]

Unit 4

Logistic Regression model and analysis: regression with a binary dependent variable, representation of the binary dependent variable, estimating the logistic regression model, assessing the goodness of fit of the estimation model, testing for significance of the coefficients, interpreting the coefficients, criteria for evaluation of logistic regression model, KS, Gini, AUC, Precision, Recall F1 score etc. Discriminant model and analysis: a two-group discriminant analysis, a three-group discriminant analysis, the decision process of discriminant analysis (objective, research design, assumptions, estimation of the model, assessing overall fit of a model, interpretation of the results, validation of the results). **[20L]**

References Books:

1. Anderson, T. W. (1984). Introduction to Multivariate Analysis, John Wiley.
2. Richard A. Johnson and Dean W. Wichern, Applied Multivariate Statistical Analysis,
3. Prentice hall India, 7th Edition, 2019.
4. Fang, K., Kotz, S., Ng K. W. (1990). Symmetric Multivariate and Related Distributions, Chapman and Hall
5. Härdle, W. K. & Simar, L. (2012). Applied Multivariate Statistical Analysis, Springer, New York
6. Härdle, W. K., Hlávka, Z. (2007). Multivariate Statistics: Exercises and Solutions,
7. Springer, New York
8. Kotz, S., Balakrishnan N. and Johnson N. L. (2000). Continuous Multivariate Distributions, Volume 1, Models and Applications, John Wiley & Sons,
9. Kshirsagar, A. M. (1983). Multivariate Analysis, Marcel Dekker
10. Morrison, D.F. (1990). Multivariate Statistical Methods, McGraw Hill Co.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. Data Science (Semester- III)
Paper Code : PSDS233
Paper : III
Title of Paper : Time Series Analysis and Forecasting
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To equip various forecasting techniques and familiarize on modern statistical methods for analyzing time series data.
2. To amalgamate the intellectual facts of the time series data to implement in the field projects scientifically.
3. To link time dependent analytical tools and building the models by extracting real time data.

Learning Outcomes

1. Students will be a clear understanding of the subject interrelated concepts and of up-to-date issues
2. Students will be getting problem-solving ability, solving social issues and industrial problems

TOPICS/CONTENTS:

Unit 1

Exploratory analysis of Time Series

Graphical display, classical decomposition model, Components and various decompositions of Time Series Models-Numerical description of Time Series: Stationarity, Auto-covariance and Autocorrelation functions, data transformations, Methods of estimation, trend, seasonal and exponential.

Smoothing Techniques

Moving Average, exponential smoothing, Holt's and Winter's methods, exponential smoothing techniques for Series with trend and seasonality, basic evaluation of exponential smoothing. **(15 L)**

Unit 2

Stationarity models:

Time series data, Trend, seasonality, cycles and residuals, Stationary, White noise processes, Autoregressive (AR), Moving Average (MA), Autoregressive and Moving Average (ARMA) and Autoregressive Integrated Moving Average (ARIMA) processes, Choice of AR and MA periods.

Non- Stationarity models:

Tests for Non-stationarity: Random walk, random walk with drift, Trend stationary, General Unit Root Tests: Dickey Fuller Test, Augmented Dickey Fuller Test. ARIMA Models: Basic formulation of the ARIMA Model and their statistical properties, Autocorrelation function (ACF), Partial autocorrelation function (PACF) and their standard Errors. **(15 L)**

Unit 3

Forecasting:

Nature of Forecasting, Forecasting methods, qualitative and quantitative methods, steps involved in stochastic model building, forecasting model evaluation. Model selection techniques: AIC, BIC and AICC – Forecasting model monitoring.

Transfer function and Intervention analysis

Transfer function models, Transfer function, noise models, Cross correlation function, Model Specification, Forecasting with Transfer function, noise models, Intervention analysis.

Unit 4

Spectral analysis

Spectral density function (s. d. f.) and its properties, s. d. f. of AR, MA and ARMA processes, Fourier transformation and period gram. **(10 L)**

Reference Books:

1. Brockwell, P.J. and Davis, R. A. *Introduction to Time Series Analysis*, Springer.
2. Chatfield, C. (2001). *Time Series Forecasting*, Chapmann & hall, London.
3. Fuller, W. A. (1996). *Introduction to Statistical Time Series*, 2nd Ed. John Wiley.
4. Hamilton N. Y. (1994). *Time Series Analysis*. Princeton University press. Princeton.
5. Kendall, Sir Maurice and Ord, J. K. (1990). Time Series (Third Edition), Edward Arnold.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. Data Science (Semester- III)
Paper Code : PSDS234
Paper : IV
Title of Paper : Artificial Intelligence
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. Students will be familiar with basic principles of AI.
2. Students will be capable of using heuristic searches.

Learning Outcome:

1. Students will be aware of knowledge-based systems.
2. Students will be able to use fuzzy logic and neural networks.
3. Students will be learn various applications domains AI.

Unit 1

Fundamentals of Artificial Intelligence: Introduction, A.I. Representation, Non-AI & AI Techniques, Representation of Knowledge, Knowledge Base Systems, State Space Search, Production Systems, Problem Characteristics, types of production systems, Intelligent Agents and Environments, concept of rationality, the nature of environments, structure of agents, problem solving agents, problem formulation. **(15)**

Unit 2

Uninformed Search Strategies: Formulation of real-world problems, Breadth First Search, Depth First Search, Depth Limited Search, Iterative Deepening Depth First Search, Bidirectional Search, Comparison of Uninformed search Strategies, Searching with partial information, Sensor-less problems, Contingency problems. **(15)**

Unit 3

Knowledge Representation Knowledge based agents, Wumpus world. Propositional Logic: Representation, Inference, Reasoning Patterns, Resolution, Forward and Backward Chaining. First order Logic: Representation, Inference, Reasoning Patterns, Resolution, Forward and

Backward Chaining. Basics of PROLOG: Representation, Structure, Backtracking. Expert System: Case study of Expert System in PROLOG. **(15 L)**

Unit 4

Introduction to Planning and ANN: Blocks world, STRIPS, Implementation using goal stack, Introduction to Neural networks: - basic, comparison of human brain and machine, biological neuron, general neuron model, activation functions, Perceptron learning rule, applications and advantages of neural networks. Brief introduction to single layer and multiplayer networks.

(15 L)

Reference Book:

1. Ivan Bratko : "Prolog Programming For Artificial Intelligence" , 2nd Edition Addison Wesley, 1440.
2. Eugene, Charniak, Drew Mcdermott: "Introduction to Artificial Intelligence.", Addison Wesley
3. Patterson: ―Introduction to AI and Expert Systems‖, PHI
4. Nilsson: ―Principles of Artificial Intelligence‖, Morgan Kaufmann.
5. Carl Townsend, ―Introduction to turbo Prolog‖, Paperback, 1483 6. Jacek M. Zurada, Introduction to artificial neural systems, Jaico Publication.
6. Elaine Rich and Kevin Knight: "Artificial Intelligence." Tata McGraw Hill
7. Stuart Russell & Peter Norvig: "Artificial Intelligence: A Modern Approach", Pearson Education, 2nd Edition.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- III)
Paper Code : PSDS235
Paper : V
Title of Paper : Text Mining and Natural Process Language
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

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

1. Describe the fundamental concepts and techniques of natural language processing.
2. Distinguish among the various techniques, taking into account the assumptions, strengths, and weaknesses of each,
3. Use appropriate descriptions, visualizations, and statistics to communicate the problems and their solutions.
4. Analyze large volume text data generated from a range of real-world applications.

Learning Outcomes:

1. Undergraduate Students: After successfully completing this course, the postgraduate students will be able to grasp the significance of natural language processing in solving real-world problems.
2. They will be able to map the appropriate processing technique to a problem and implement the technique.
3. They will also be able to demonstrate required design skills for large collection sets.
4. They will be able to comprehend the state-of-the-art advanced nlp research articles and present them to an audience.
5. They will also be able to propose extension of existing nlp techniques for solving a range of problems.

Unit 1 **(15 L)**

Introduction to text data, structure of text data, Working with Text Data, Character Encodings, Tokenization, Parsing, Stemming, APIs, Web Scraping, Regular Expressions, Spelling Correction, representation of the unstructured text documents with appropriate format and structure to support later automated text mining algorithms.

Unit 2 **(15 L)**

Probabilistic models for text mining: Naïve Bayes, basic supervised text categorization algorithms: k Nearest Neighbor (kNN) and Logistic Regression, Support Vector Machines and Decision Trees.

Unit 3 **(15 L)**

Text clustering: introduction, typical types of clustering algorithms: connectivity-based clustering (hierarchical clustering) and centroid-based clustering (e.g., k-means clustering).

Unit 4 **(15 L)**

Sentiment Analysis: Introduction of sentiment analysis, task of extracting subjective information in source materials, problems in sentiment analysis: sentiment polarity prediction, review mining, and aspect identification. Social media and network analysis: characteristic of social network (inter-connectivity, and introduce Google's winning algorithm Page Rank), social influence analysis and social media analysis.

References Books:

1. Mining Text Data. Charu C. Aggarwal and Cheng Xiang Zhai, Springer, 2012.
2. Speech & Language Processing. Dan Jurafsky and James H Martin, Pearson Education India, 2000.
3. Introduction to Information Retrieval. Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schuetze, Cambridge University Press, 2007.
4. Foundations of Statistical Natural Language Processing by Christopher Manning and Hinrich Schütze.
5. Natural Language Processing with Pythonby Steven Bird, Ewan Klein and Edward Loper.
6. Survey of Text Mining Clustering, Classification, and Retrieval by Michael W. Berry
7. Creating Value With Social Media Analytics Managing, Aligning, and Mining Social Media Text, Networks, Actions, Location, Apps, Hyperlinks, Multimedia, & Search Engines Data by Gohar F. Khan.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- III)
Paper Code : PSDS236
Paper : VI
Title of Paper : Data Visualization Using Tableau
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. The main focus of Tableau software is for you to better understand your datasets, especially large datasets.
2. Able to handle 'big' data.
3. To implement the best design practices, and use the most appropriate chart for a particular situation.
4. To build interactive Tableau dashboards and construct a data story using Tableau Story point

Learning Outcomes:

1. Build data visualizations in Tableau
2. Use data hierarchies, filters, groups, sets, and calculated fields.
3. Create map-based data visualizations in Tableau.

TOPICS/CONTENTS:

1. **Understanding Data:** What is data, Foundations for building Data Visualizations, Getting started with Tableau Software, Using Data file formats, Connecting Data to Tableau, Creating basic charts (line, bar charts, Treemaps), Using the Show me panel.
2. **Tableau Calculations:** Overview of SUM, AVR, and Aggregate features, Creating custom calculations and fields, Applying new data calculations to your visualization.
3. **Formatting Visualizations:** Formatting Tools and Menus, Formatting specific parts of the view, Editing and Formatting Axes.
4. **Manipulating Data in Tableau:** Cleaning-up the data with the Data Interpreter, Structuring your data, Sorting and filtering Tableau data, Pivoting Tableau data.
5. **Advanced Visualization Tools:** Using Filters, Using the Detail panel, Using the Size panels, Customizing filters, Using and Customizing tooltips, Formatting your data with colours.
6. **Creating Dashboards & Stories:** Using Storytelling, Creating dashboard and Story, Design for different displays, Adding interactivity to Dashboard.
7. **Distributing & Publishing Visualization:** Tableau file types, Publishing to Tableau Online, Sharing visualization.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- IV)
Paper Code : PSDS241
Paper : I
Title of Paper : Machine Learning
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1 Understanding the nature of problems solved with Machine Learning.

2 To study different supervised learning algorithms and unsupervised learning algorithms.

3 To understand the application development process using ML.

Learning Outcomes:

Students are expected to

1. Have a good understanding of the fundamental issues and challenges of machine learning: data, model selection, model complexity, etc.
2. compare the strengths and weaknesses of many popular machine learning approaches
3. design and implement various machine learning algorithms in a range of real-world Applications.

TOPICS/CONTENTS:

Unit 1

Introduction to Data and Machine Learning

Interpretability of Machine Learning Models - Why interpretability is necessary, Model agnostics methods of interpretability – Partial Dependence Plot (PDP), Individual Conditional Expectation (ICE), Shapley values etc. Essentials of Data and its analysis, Framework of Data Analysis, History of Machine Learning, Machine Learning Vs Statistical Learning, Types of Machine Learning Algorithms, Supervised Learning, Unsupervised Learning, Reinforcement Learning. **(15L)**

Unit 2

Understanding Regression Analysis

Linear Regression, Multiple Regression, Logistic Regression

Classification Techniques

k-nearest neighbor, decision tree, Naïve Bayesian, Support vector machine ,artificial neural network, convoluted neural network (CNN),Classification based on logistic regression. **(15L)**

Unit 3 Clustering

K means clustering, Association Rule Mining, Apriori Algorithm

Model evaluation and selection methods:

Metrics for evaluating classifier performance (confusion matrix), holdout method and random sampling, cross validation, bootstrap, ROC curves, AIC, BIC, CIC, DIC (information criterion), bias variance tradeoff. **(15L)**

Unit 4

Techniques to improve classification accuracy

Bagging, boosting, Ada boosting, Random forest, gradiant boosting. Self-Organizing Map (SOM), EM algorithm, market basket analysis, text mining: sentiment analysis, word frequency analysis, N-grams and correlation, topic modeling. **(15L)**

References Books:

1 Jiawei Han, Micheline Kamber, Jian Pei, Data Mining: Concepts and Techniques, 3rd Edition.

2 Margaret H. Dunham, S. Sridhar, Data Mining - Introductory and Advanced Topics, Pearson Education

3 Tom Mitchell, Machine Learning‖, McGraw-Hill, 1997

4 R.O. Duda, P.E. Hart, D.G. Stork., Pattern Classification, Second edition. John Wiley and Sons, 2000.

5 Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer 2006

6 Ian H. Witten, Data Mining: Practical Machine Learning Tools and Techniques, Eibe Frank Elsevier /(Morgan Kauffman)

7 Bing Liu: Web Data Mining: Exploring Hyperlinks, Contents and Usage Data, Springer (2006).

8 Soumen Chakrabarti: Mining the Web: Discovering knowledge from hypertext data, Elsevier (2003).

9 Christopher D Manning, Prabhakar Raghavan and HinrichSchütze: An Introduction to Information Retrieval, Cambridge University Press (2009)

SYLLABUS (CBCS) FOR M.Sc. Data Science
(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- IV)
Paper Code : PSDS242
Paper : II
Title of Paper : Discrete Data Analysis
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To familiarize students to learn about regression analysis and model building for discrete data.
2. To investigate possible diagnostics in various techniques for discrete data.
3. To acquaint students with validation of the generalized linear models using hypothesis testing and analysis of deviance.

Learning Outcomes:

1. Students can estimate the parameters and fit different models to discrete data.
2. Students can apply linear as well as nonlinear models to analyze discrete data.
3. Students will be able to validate the model such as Poisson regression and Logistic Regression using cross validation techniques.

TOPICS/CONTENTS:

Unit 1

Log linear model for two and three dimensional contingency tables: Interpretation of parameters, comparison with ANOVA and regression. ML estimation of parameters, likelihood ratio tests for various hypotheses including independence. Marginal and conditional independence, partial association, models with quantitative levels. **(15 L)**

Unit 2

Generalized linear models: concept of generalized linear model, Link function, ML estimation, large sample tests about parameters, goodness of fit, analysis of deviance, introduction to Poisson regression. **(15 L)**

Unit 3

Poisson regression: ML and Quasi-likelihood estimation of parameters, testing significance of coefficients, goodness of fit, power family of link functions, over dispersion: Types, causes and remedies. Negative Binomial regression: NB-2 model. **(15 L)**

Unit 4

Non-parametric regression and Interpolating and smoothing splines for simple regression. Use of cross-validation applications to logistic and Poisson regression. **(15 L)**

References Books:

1. Yvonne M. Bishop, Stephen E. Fienberg, Paul W. Holland Discrete (1975): Multivariate Analysis: Theory and Practice
2. Hosmer D.W. and Lemeshow S. (2000): Applied Logistic Regression, 2nd Ed. Wiley, New York.
3. Agesti A. (1990) : Categorical Data Analysis. Wiley, New York.
4. R. Christensen (1997): Log-Linear Models and Logistic Regression. 2nd Ed. Springer, New York.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- IV)
Paper Code : PSDS243
Paper : III
Title of Paper : Supply Chain and Logistics
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

1. To develop an understanding of the various supply chain and logistics processes.
2. To develop knowledge on structures, decision phases, measures and tools of supply chains.
3. To develop understanding on the strategic, tactical and operational decision tools of supply chains.
4. To impart knowledge on logistics management and related advanced tools and techniques.
5. To develop and analyse the role of digital transformation of the supply chains and logistics.

Learning Outcomes:

The students will

1. Understand the structures, decision phases, measures and tools of supply chains.
2. Understand the strategic, tactical and operational decision tools of supply chains.
3. Understand knowledge on logistics management and related advanced tools and techniques.

TOPICS/CONTENTS:

Unit 1

Introduction to Supply Chain: The basic Supply Chain model, Generalized Supply Chain Model, Value Chain and Value Chain Analysis, Supply Chain Effectiveness: Strategy, Metrics, Technology, Supplier Performance, Integration and Collaboration, Risk Mitigation, Supply Chain Applications, Information Functionality – The Supply Chain. **[15L]**

Unit 2

Supply Chain Design and Planning: Supply Chain Configuration, Extent of Vertical Integration, Outsourcing and Offshoring, Location Decisions, Capacity Planning.

Planning Demand & Supply: Planning demand and supply in supply chains – Forecasting techniques for supply chains, Seasonal Forecasting Models, Measure of Forecast errors.

[15L]

Unit 3

Logistics in 21st Century, Definition and Concept of Logistics, Concept of Logistics, Logistics Value Proposition, Service Benefits, Cost Minimization, Logistics Value Generation, The Work of Logistics: Order Processing, Inventory, Transportation, Warehousing, Materials Handling, and Packaging, Facility Network Design, Logistical Operations: Inventory Flow, Physical Distribution, Manufacturing Support, Procurement ,Information Flow, Planning and Coordination Flows, Operational Requirements, Logistical Operating Arrangements : Echelon Structured Logistics, Direct Structured Logistics, Flexible Logistics System, Principles of Logistics Information. **[20L]**

Unit 4

Logistics Management: 3PL, 4PL, Design Options for Transportation Network. Routing, scheduling and Sequencing in Transportation, Vehicle Routing Problems. Quantitative Examples. Reverse Logistics: Reverse logistics and Closed Loop Supply Chains. Advanced Logistics Decision Models: Bin Packing Problems, Fixed Charge Problems, Knapsack Problems, Multi-stage transportation problems. **[10L]**

References Books:

1. Neha Tikoo LOGISTICS AND SUPPLY CHAINMANAGEMENT
2. Dr. Dawei Lu Fundamentals of Supply Chain Management
3. David Simchi – Levi & Philip Kaminsk, Designing and Managing the Supply Chain, McGraw-Hill Companies Inc.
4. David Taylor and David Brunt, Manufacturing Operations and Supply Chain Management, Vikas Thomson Learning, 2001.
5. Donald J. Bowersox & David J. Closs, Logistical Management, TMH.
6. Jeremy F. Shapiro, Modeling and Supply Chain,. Thomson Learning, 2001.

SYLLABUS (CBCS) FOR M.Sc. Data Science

(With effect from Academic Year 2022-2023)

Class : M. Sc. (Semester- IV)
Paper Code : PSDS244
Paper : IV
Title of Paper : Deep Learning
Credit : 4 credits
No. of lectures : 60

Learning Objectives:

Deep learning eliminates some of data pre-processing that is typically involved with machine learning. These algorithms can ingest and process unstructured data, like text and images, and it automates feature extraction, removing some of the dependency on human experts.

Learning Outcome:

1. Evaluate, in the context of a case study, the advantages and disadvantages of deep learning neural network architectures and other approaches.
2. Implement deep learning models in Python using the PyTorch library and train them with real-world datasets.
3. Design convolution networks for handwriting and object classification from images or video.
4. Design recurrent neural networks with attention mechanisms for natural language classification, generation, and translation.
5. Evaluate the performance of different deep learning models (e.g., with respect to the bias-variance trade-off, over fitting and under fitting, estimation of test error).
6. Perform regularization, training optimization, and hyper parameter selection on deep models.

Unit 1:

Deep learning basics: Introduction, History, capabilities, the perceptron, Neural network learning: Back-Propagation, Practical network training o Autoencoders, Batch-normalization, why does it work? Over fitting and generalization. **[15 L]**

Unit 2:

Convolutional neural networks: Introduction to CNNs, Convolution, Correlation, Filtering. CNN architectures, Detection and Segmentation, Visualizing and Understanding, Advanced CNNs for computer vision. **[15L]**

Unit 3:

Advanced Deep architectures: Recurrent Neural networks (RNNs), Advanced RNN: LSTM, GRU, Generative Adversarial Networks (GANs), Advanced GANs, Advanced topics, Recent papers, Influential papers, Deep reinforcement learning, Deep Learning: Good -> Great, Visual Question Answering, Visual Dialog, Novel deep methods (Deep internal learning, Deep image prior). **[20 L]**

Unit 4:

Tools: Tensor flow, Pytorch, Practical sessions: Computer Vision, Sequence modeling, Natural / Biological signals **[10 L]**

References Books:

1. Deep Learning with Python by François Chollet, Manning Publications Co, ISBN: 9781617294433
2. Deep Learning - A Practical Approach by Rajiv Chopra, Khanna Publications, ISBN: 9789386173416
3. Deep Learning by Ian Good fellow and Yoshua Bengio and Aaron Courville Published by An MIT Press book.