DISCIPLINE SPECIFIC ELECTIVE COURSE-5(iii): MATHEMATICAL DATA SCIENCE

CREDIT DISTRIBUTION, ELIGIBILITY AND PRE-REQUISITES OF THE COURSE

Course title & Code	Credits	Credit distribution of the course			Eligibility criteria	Pre-requisite of the course (if any)
		Lecture	Tutorial	Practical/ Practice		
Mathematical Data Science	4	3	0	1	Class XII pass with Mathematics	Basic knowledge of R/Python, Probability and Statistics

Learning Objectives: The main objective of this course is to:

- Introduce various types of data and their sources, along with steps involved in data science case-study, including problems with data and their rectification and creation methods.
- Cover dimensionality reduction techniques, clustering algorithms and classification methods.

Learning Outcomes: The course will enable the students to:

- Gain a comprehensive understanding of data science, its mathematical foundations including practical applications of regression, principal component analysis, singular value decomposition, clustering, support vector machines, and *k*-NN classifiers.
- Demonstrate data analysis and exploration, linear regression techniques such as simple, multiple explanatory variables, cross-validation and regularization using R/Python.
- Use real-world datasets to practice dimensionality reduction techniques such as PCA, SVD, and multidimensional scaling using R/Python.

SYLLABUS OF DSE-5(iii)

UNIT-I: Principles of Data Science (12 hours)

Types of Data: nominal, ordinal, interval, and ratio; Steps involved in data science case-study: question, procurement, exploration, modeling, and presentation; Structured and unstructured data: streams, frames, series, survey results, scale and source of data – fixed, variable, high velocity, exact and implied/inferred; Overview of problems with data – dirty and missing data in tabular formats – CSV, data frames in R/Pandas, anomaly detection, assessing data quality, rectification and creation methods, data hygiene, meta-data for inline data-description-markups such as XML and JSON; Overview of other data-source formats – SQL, pdf, Yaml, HDF5, and Vaex.

Unit-II: Mathematical Foundations (15 hours)

Model driven data in R^n, Log-likelihoods and MLE, Chebyshev, and Chernoff-Hoeffding inequalities with examples, Importance sampling; Norms in Vector Spaces– Euclidean, and metric choices; Types of distances: Manhattan, Hamming, Mahalanobis, Cosine and angular distances, KL divergence; Distances applied to sets– Jaccard, and edit distances; Modeling text with distances; Linear Regression: Simple, multiple explanatory variables, polynomial, cross-validation, regularized, Lasso, and matching pursuit; Gradient descent.

Unit-III: Dimensionality Reduction, Clustering and Classification (18 hours)

Problem of dimensionality, Principal component analysis, Singular value decomposition (SVD), Best *k*-rank approximation of a matrix, Eigenvector and eigenvalues relation to SVD, Multidimensional scaling, Linear discriminant analysis; Clustering: Voronoi diagrams, Delaunay triangulation, Gonzalez's algorithm for *k*-center clustering, Lloyd's algorithm for *k*-means clustering, Mixture of Gaussians, Hierarchical clustering, Density-based clustering and outliers, Mean shift clustering; Classification: Linear classifiers, Perceptron algorithm, Kernels, Support vector machines, and *k*-nearest neighbors (*k*-NN) classifiers.

Essential Readings

1. Mertz, David. (2021). Cleaning Data for Effective Data Science, Packt Publishing.
2. Ozdemir, Sinan. (2016). Principles of Data Science, Packt Publishing.
3. Phillips, Jeff M. (2021). Mathematical Foundations for Data Analysis, Springer. (https://mathfordata.github.io/).

Suggestive Readings

- Frank Emmert-Streib, et al. (2022). Mathematical Foundations of Data Science Using R. (2nd ed.). De Gruyter Oldenbourg.
- Wes McKinney. (2022). Python for Data Analysis (3rd ed.). O'Reilly.
- Wickham, Hadley, et al. (2023). R for Data Science (2nd ed.). O'Reilly.

Practical (30 hours)- Practical work to be performed in Computer Lab using R/Python:

1. To explore different types data (nominal, ordinal, interval, ratio) and identify their properties.
2. To deal with dirty and missing data, such as imputation, deletion, and data normalization.
3. Use the real-world datasets (https://data.gov.in/) to demonstrate the following:
   e) Data analysis and exploration, linear regression techniques such as simple, multiple explanatory variables, cross-validation, and regularization.
   f) Dimensionality reduction techniques such as principal component analysis, singular value decomposition (SVD), and multidimensional scaling.
   g) Clustering algorithms such as *k*-means, hierarchical, and density-based clustering and evaluate the quality of the clustering results.
   h) Classification methods such as linear classifiers, support vector machines (SVM), and *k*-nearest neighbors (*k*-NN).