17HS064 Foundations of Data Science

Course Objectives

Modern scientific, engineering, and business applications are increasingly dependent on data, existing traditional data analysis technologies were not designed for the complexity of the modern world. Data Science has emerged as a new, exciting, and fast-paced discipline that explores novel statistical, algorithmic, and implementation challenges that emerge in processing, storing, and extracting knowledge from Big Data.

Course Outcomes

1. Able to apply fundamental algorithmic ideas to process data.

2. Learn to apply hypotheses and data into actionable predictions.

3. Document and transfer the results and effectively communicate the findings using visualization techniques.

UNIT I

INTRODUCTION TO DATA SCIENCE : Data science process – roles, stages in data science project – working with data from files – working with relational databases – exploring data – managing data – cleaning and sampling for modeling and validation – introduction to NoSQL.

UNIT II

MODELING METHODS : Choosing and evaluating models – mapping problems to machine learning, evaluating clustering models, validating models – cluster analysis – K-means algorithm, Naïve Bayes – Memorization Methods – Linear and logistic regression – unsupervised methods.

UNIT III

INTRODUCTION TO R Language: Reading and getting data into R – ordered and unordered factors – arrays and matrices – lists and data frames – reading data from files – probability distributions – statistical models in R - manipulating objects – data distribution.

UNIT IV

MAP REDUCE: Introduction – distributed file system – algorithms using map reduce, Matrix-Vector Multiplication by Map Reduce – Hadoop - Understanding the Map Reduce architecture - Writing Hadoop Map Reduce Programs - Loading data into HDFS - Executing the Map phase - Shuffling and sorting - Reducing phase execution.

UNIT V

DELIVERING RESULTS : Documentation and deployment – producing effective presentations– Introduction to graphical analysis – plot() function – displaying multivariate data – matrix plots – multiple plots in one window - exporting graph - using graphics parameters. Case studies.

Reference Books

1. Nina Zumel, John Mount, “Practical Data Science with R”, Manning Publications, 2014.
2. Jure Leskovec, Anand Rajaraman, Jeffrey D.Ullman, “Mining of Massive Datasets”, Cambridge University Press, 2014.
3 .Mark Gardener, “Beginning R - The Statistical Programming Language”, John Wiley & Sons, Inc., 2012.
4. W. N. Venables, D. M. Smith and the R Core Team, “An Introduction to R”, 2013.
5. Tony Ojeda, Sean Patrick Murphy, Benjamin Bengfort, Abhijit Dasgupta, “Practical Data Science Cookbook”, Packt Publishing Ltd., 2014.
6. Nathan Yau, “Visualize This: The FlowingData Guide to Design, Visualization, and Statistics”, Wiley, 2011.
7 .Boris lublinsky, Kevin t. Smith, Alexey Yakubovich, “Professional Hadoop Solutions”, Wiley, ISBN: 9788126551071, 2015.

Student Activity:

1. Collect data from any real time system and create clusters using any clustering algorithm
2. Read the student exam data in R perform statistical analysis on data and print results.