CSE 327: Introduction to Data Science

Programme: B.Tech. (CSE) Year: 3 Semester : V
Course : Core Course Credits : 3 Hours : 40

Course Context and Overview:

Availability of huge data and affordable computing is increasing the desire to get better information out of data which can be structured and/or unstructured. Extraction of this information requires interdisciplinary knowledge involving statistics, data analysis, machine learning and related methods. This course includes concepts largely from statistics, machine learning and data mining to enable the students to analyze data and extract information out of it.

Prerequisite Courses:

Design and Analysis of Algorithms, Probability and Statistics

Course Outcomes (COs):

On completion of this course, the students will have the ability to:
CO1 Do data pre-processing and explore the given data set
CO2 Analyse the given data set by using various techniques for both numerical and categorical data
CO3 Apply various machine learning algorithms for prediction, forecasting and other related problems
CO4 Handle text data, do pre-processing, POS tagging and word sense disambiguation

Course Topics:

Contents	**Lecture Hours**	
UNIT 1 **Introduction**		2
Course Overview and Description, Motivation with some examples, Objectives	2	
UNIT 2 **Statistical Analysis**		10

Descriptive Statistics and Exploratory Data Analysis: Graphical Approaches, Measures of Location, Measures of Spread, Random Variables and Probability Distributions	4	
Inferential Statistics: Motivation, Estimating unknown parameters, Testing Statistical Hypothesis, One sample and two small tests, Regression and ANOVA and test of independence	6	
UNIT 3 **Data Preprocessing**		3
Data cleaning, Data Reduction, Data Transformation, Data Discretization, Similarity & Dissimilarity measures	3	
UNIT 4 **Introduction to Machine Learning**		17
Supervised and Unsupervised Learning, Algorithmic frameworks vs. Model based frameworks	1	
Supervised Learning: Decision Tree, Bayes rule, Naïve Bayes Classifier, K-NN Classifier, Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Ensemble Methods	9	
Unsupervised Learning: Cluster Analysis, Partition Methods, Hierarchical Methods	5	
Evaluation Methodology: Experimental Setup, Measuring Performance of Models, Interpretation of Results	2	
UNIT 5 **Text Analytics and Data Science Pipeline**		8
Basics on Text Analysis: Words and Tokens, HMM POS tagging – The Viterbi Algorithm, Word Sense Disambiguation, Basic Text similarity measures	4	
Data Science Pipeline (with two domain-specific case studies for the following topics): Data Collection, Data Preprocessing, Data Exploration, Data Modeling and Data Interpretation	4	

Textbook references: No Textbooks for this course.

Reference books:

Tom Mitchell. *Machine Learning*. 1st edition, McGraw Hill, 1997**.**
P-N Tan, M Steinbach, A. Karpatne, V Kumar. *Introduction to Data Mining*. 2nd edition, Pearson Education, 2018.
Sheldon M Ross. *Introduction to Probability and Statistics*. 3rd edition, Elsevier, 2004.
D Montgomery, GC Runger. *Applied Statistics and Probability for Engineers*. 5th edition, John Wiley and Sons, 2010.
Daniel Jurafsky and James H. Martin. *Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, Pearson, 2nd edition, 2014.

Evaluation Methods:

Component	***Weightage (%)***
Assignment 1	5%
Assignment 2	5%
Quiz	10%
Project	20%
Mid Term	25%
End Term	35%

Prepared By: Subrat K Dash, Sakthi Balan Muthiah
First Update: 30th July 2018
Second Update (by Dr. Sakthi): 7th July 2020 (After discussion with Dr. Subrat, Dr. Vibhor and Dr. Sudheer)
Third Update (Only Evaluation is updated): 3rd Aug 2020