CSE 327: Introduction to Data Science

Programme: B.Tech. (CSE) Year: III 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

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 Description, Course overview	

UNIT 2 **Statistical Analysis**	10
Descriptive Statistics and Exploratory Data Analysis: Graphical Approaches, Measures of Location, Measures of Spread, Random Variables and Probability Distributions	
Inferential Statistics: Motivation, Estimating unknown parameters, Testing Statistical Hypothesis, One sample and two small tests, Regression and ANOVA and test of independence	
UNIT 3 **Data Preprocessing**	3
Data cleaning, Data Reduction, Data Transformation, Data Discretization, Similarity & Dissimilarity measures	
UNIT 4 **Introduction to Machine Learning**	17
Supervised and Unsupervised Learning, Algorithmic frameworks vs. Model based frameworks	
Supervised Learning: Bayes rule, Naïve Bayes Classifier, K-NN Classifier, Logistic Regression, Linear Discriminant Analysis, Support Vector Machines, Ensemble Methods	
Unsupervised Learning: Cluster Analysis, Partition Methods, Hierarchical Methods	
Evaluation Methodology: Experimental Setup, Measuring Performance of Models, Interpretation of Results	
UNIT 5 **Case Studies and Advanced Learning Models**	8
Basics on Text Analysis: POS tagging, Word sense disambiguation, Text similarity measures, Intro to WordNet	

Reinforcement Learning: Model-based learning, Passive RE, Model-Free learning, Active RE, Exploration versus Exploitation, Q-Learning

Textbook references:
Text Book:

Reference books:

- Tom Mitchell. Machine Learning. McGraw Hill, 1997.
- P-N Tan, M Steinbach, V Kumar. Introduction to Data Mining. Pearson Education, 2016.
- Introduction to Probability and Statistics by Sheldon M Ross, 3rd edition, Elsevier, 2004 .
- Applied Statistics by Montgomery. 3rd edition, John Wiley and Sons
- Artificial Intelligence: Modern Approach by Russell and Norvig (2nd edition)

Evaluation Methods:

Item	***Weightage***
Continuous evaluation (quiz, assignment, class test etc.)	20%
Project	15%
Midterm	25%
Final Examination	40%

Prepared By: Subrat K Dash, Sakthi Balan Muthiah
Last Update: 30th July 2018