CMUA22201: Data Science and Machine Learning

Teaching Scheme	Examination Scheme					
Credits: 04 **Lecture (L):** 03 hrs / week **Practical (T):** 02 hrs / week	CIE	ISE	SCE	ESE	TW	Total
	20	20	20	40	25	125

Prerequisites : Probability and Statistics

Course Objective(s):

1. Understand the data science life cycle
2. Learn the statistical methods data pre-processing
3. Learn and apply unsupervised approach for prediction.
4. Learn and apply Supervised models for prediction
5. Interpret classification outcome
6. Learn effective data visualization

Course Outcomes:

After completion of the course, student will be able to:

1. Describe the Data Science Process and explore components interaction.
2. Apply statistical methods for pre-processing and extracting meaning from data to the application dataset.
3. Apply specific supervised regression machine learning algorithm for a particular problem.
4. Apply specific supervised-classification machine learning algorithm for a particular problem.
5. Apply specific Unsupervised machine learning algorithm for a particular problem
6. Analyse the outcome of an algorithm in terms of efficiency.

Contents

Unit I: Introduction to Data Science (6 Hrs)

Introduction: Big data overview, state of the practice in Analytics- BI Vs Data Science, Current Analytical Architecture, drivers of Big Data, Emerging Big Data Ecosystem and new approach. Philosophy of Exploratory Data Analysis, The Data Science Process, A Data Scientist's Role Data Analytic Life Cycle: Overview, phase 1- Discovery, Phase 2- Data preparation, Phase 3- Model Planning, Phase 4- Model Building, Phase 5- Communicate Results, Phase 6-Operationalize. Case Study. Statistical description and inference of Data.

Unit II: Introduction to Machine Learning (7 Hrs)

Introduction to Machine Learning, Applications, Introduction to Machine Learning Techniques: Supervised Learning, Unsupervised Learning and Reinforcement Learning, Data formats, Creating training and testing datasets. Preprocessing: Data Cleaning, Data Transformation and Data reduction.

Unit III: Supervised Models I (7 Hrs)

Regression: Linear Regression, Multiple Regression, Polynomial regression, Logistic Regression, K-Nearest neighbor, Support Vector Machines.

Unit IV: Supervised Models II (8 Hrs)

Classification Decision trees- Overview, decision tree algorithm, evaluating a decision tree using Gini Index and Entropy, Random forest, Naïve Bayes – Bayes Theorem and Algorithm, Naïve Bayes Classifier, smoothing, diagnostics. Diagnostics of classifiers, additional classification methods.

Unit V: Unsupervised Modelling (7 Hrs)

Cluster Analysis: Basic Concepts and Methods, Partitioning Methods: k-Means: A Centroid Based Technique, k-Medoids: A Representative Object-Based Technique, Hierarchical Methods: Agglomerative versus Divisive Hierarchical Clustering.

Unit VI: Model Evaluation and Selection (7 Hrs)

Metrics for Evaluating Classifier Performance Model Selection Using Statistical Tests of Significance Comparing Classifiers Based on Cost–Benefit and ROC Curves, Confusion Matrix, F-Measure, Precision, Recall. Cross validation, Underfitting and Overfitting, Bias and Variance, Regularization, Ridge regression, Lasso and ElasticNet regression

Text-Books:

1. Data Science and Machine Learning: Mathematical and Statistical Methods By D.P. Kroese, Z.I. Botev, T. Taimre, R. Vaisman, *Chapman and Hall/CRC, Boca Raton, 2019*.
2. Machine Learning with Python Cookbook Practical Solutions from Preprocessing to Deep Learning by Chris Albon, O'Reilly
3. Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems by Aurélien Géron, O'Reilly

Reference Books :

1. Introducing Data Science, Davy Cielen, Aron D.B Meysman. MANNING publishing Data Science and Machine Learning, Publisher: Sigma Data Systems, United States, ISBN: 978-1655848049

2. Introduction to Machine Learning with Python A Guide for Data Scientists, Andreas C. Müller and Sarah Guido, O'Reilly.

List Of Assignments:

1. Perform the following operations using Python on suitable data sets, read data from different formats(like csv, xls),indexing and selecting data, sort data, describe attributes of data, checking data types of each column, counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa), identifying missing values and fill in the missing values.
2. Perform the following operations using Python on the data sets Compute and display summary statistics for each feature available in the dataset. (e.g. minimum value, maximum value, mean, range, standard deviation, variance and percentiles), Create a histogram for each feature in the dataset to illustrate the feature distributions.
3. Implement following algorithms using Python on suitable data sets.
   i. Linear Regression
   ii. Polynomial Regression
   iii. Logistic Regression
4. Implement following algorithms using Python on suitable data sets.
   i. K-Nearest neighbour
   ii. Support Vector Machines
5. Implement following algorithms using Python on suitable data sets.
   i. Decision Tree
   ii. Naïve Bayes
   iii. Random Forest
6. Implement following algorithms using Python on suitable data sets.

i. K-means
ii. k-Medoids

List of MOOC / NPTEL Courses:

NPTEL Course “Introduction to Machine Learning”, Prof Balaraman Ranvindran, IIT Madras
https://nptel.ac.in/courses/106106139
NPTEL Course” Machine Learning” ML By Prof. Carl Gustaf Jansson , KTH, The Royal Institute of Technology
https://onlinecourses.nptel.ac.in/noc24_cs60/preview
NPTEL Course “Introduction To Machine Learning “, Prof. Sudeshna Sarkar, IIT Kharagpur