Bansilal Ramnath Agarwal Charitable Trust’s

Vishwakarma Institute of Information Technology, Pune-48

(An Autonomous Institute affiliated to Savitribai Phule Pune University)

Department of Computer Science and Engineering (Data Science)

CD22231: MACHINE LEARNING

Teaching Scheme	Examination Scheme
Credits : 4 Lectures : 3 Hrs/week Practical : 2 Hr/week	Continuous Internal Evaluation (CIE): 20 Marks Skills & Competency Exam (SCE): 20 Marks End Semester Examination (ESE): 40 Marks PR: 20 Marks

Prerequisites: Probability and Statistics

Course Objective(s):

1. To introduce basics of Machine Learning.
2. To learn the statistical methods data pre-processing
3. To learn and apply unsupervised approach for prediction.
4. To learn and apply Supervised models for prediction
5. To interpret classification outcome
6. To 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 unsupervised machine learning algorithm for a particular problem.
4. Apply specific supervised machine learning algorithm for a particular problem.
5. Analyze the outcome in terms of efficiency.
6. Analyze and organize data using visualization tools.

Contents

Unit I: Introduction to Machine Learning (6 Hrs)

Introduction to Machine Learning:ML vs AI vs Data Science, Real life applications.

Types of Learning: Supervised Learning Unsupervised Learning, Semi-Supervised Learning, Reinforcement Learning

Models of Machine learning: Geometric model, Probabilistic Models, Logical Models, Grouping models, Parametric and non-parametric models

Unit II: Pre-processing and Extracting meaning from Data (6 Hrs)

Types of Data: Qualitative and Quantitative data.

Concept of Feature: Feature construction, Feature Selection and Transformation, Dataset Preparation: Training Vs. Testing Dataset.

Dataset Preparation: Training Vs. Testing Dataset, Identifying Missing values and approaches, Noisy Data Extraction, Data Cleaning as a process, Data reduction, Data Transformation and Discretization: Data Transformation by Normalization, Discretization by Binning Discretization by Histogram Analysis Discretization by Cluster.
Dataset Validation Techniques: Hold-out, k-fold Cross validation, Leave-One-Out Cross-Validation (LOOCV).
Dimensionality Reduction: Introduction, Subset Selection, Principal Components Analysis, Factor Analysis, Multidimensional Scaling, Linear Discriminant Analysis

Unit III: Model Evaluation and Selection (6 Hrs)

Binary Classification: Metrics for Evaluating Classifier Performance - Confusion Matrix, Accuracy, Precision, Recall, ROC Curves, AUC, F-Measure

Multi-class Classification: Metrics for Evaluating Classifier Performance - Per-class Precision and Per-Class Recall, weighted average precision and recall -with example, Handling more than two classes

Unit IV: Supervised Models– I (6 Hrs)

Regression: Linear Regression, Multiple Regression, Logistic Regression
Methods: Ridge Regression, Lasso Regression, Least Angle Regression,
Cost Functions: MSE, MAE, R-Square, Optimization of Simple Linear Regression with Gradient Descent, Estimating the values of the regression coefficients
SVM: Introduction to SVM, Computing Support Vector Classifier for classification, Soft Margin SVM, Introduction to various SVM Kernel to handle non-linear data – RBF, Gaussian, Polynomial, Sigmoid.
KNN: Introduction to K-Nearest neighbour

Unit V: Supervised Models - II(6 Hrs)

Decision trees: Overview, decision tree algorithm, evaluating a decision tree using Gini Index and Entropy,
Naive Bayes: Bayes Theorem, Naïve Bayes Classifier, smoothing, diagnostics. Diagnostics of classifiers

Unit VI: Unsupervised Models (6 Hrs)

Cluster Analysis: Basic Concepts and Methods, Proximity Matrices
Clustering Algorithms: K-mean, Gaussian Mixtures as Soft K-means Clustering,
Partitioning Methods: k-Means, A Centroid Based Methods: k-Medoids, Hierarchical Methods: Agglomerative versus Divisive Hierarchical Clustering, Dendrogram for hierarchical clustering

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.*

Reference Books:

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

List of MOOC / NPTEL Courses:

1. Introduction to Machine Learning: https://nptel.ac.in/courses/106/106/106106139/
2. Machine Learning: https://nptel.ac.in/courses/106/106/106106202/
3. Machine Learning for Science and Engineering applications: https://nptel.ac.in/courses/106/106/106106198/
4. Introduction to Machine Learning:
5. https://nptel.ac.in/courses/106/105/106105152/

List Of Assignments:

Implement any 6 assignments out of 8

1. Perform the following operations using R/Python on suitable data sets:
   a) read data from different formats (like csv, xls)
   b) Find Shape of Data
   c) Find Missing Values
   d) Find data type of each column
   e) Finding out Zero's
   f) Indexing and selecting data, sort data,
   g) Describe attributes of data, checking data types of each column,
   h) counting unique values of data, format of each column, converting variable data type (e.g. from long to short, vice versa),

2. Perform the following operations using R/Python on the data sets:
   a) 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
   b) Illustrate the feature distributions using histogram.
   c) Data cleaning, Data integration, Data transformation, Data model building (e.g. Classification)

3. Visualize the data using R/Python by plotting the graphs for assignment no. 1 and 2. Consider suitable data set. Use Scatter plot, Bar plot, Box plot, Pie chart, Line Chart.

4. Apply appropriate ML algorithm on a dataset collected in a cosmetics shop showing details of customers to predict customer response for special offer.
   Create confusion matrix based on above data and find
   a) Accuracy

b) Precision
c) Recall
d) F-1 score

5. Write a program to do following:
Data Set : students can get dataet related to income and expenditure .
This dataset givesthe data of Income and money spent by the customers visiting a shopping mall.
The data set contains Customer ID, Gender, Age, Annual Income, Spending Score. Therefore, asa mall owner you need to find the group of people who are the profitable customers for the mallowner. Apply at least two clustering algorithms (based on Spending Score) to find the group ofcustomers.

a) Apply Data pre-processing
b) Perform data-preparation (Train-Test Split)
c) Apply Machine Learning Algorithm
d) Evaluate Model.
e) Apply Cross-Validation and Evaluate Mode

6. Assignment on Regression technique.
Download temperature data from below link.
https://www.kaggle.com/venky73/temperaturesof-india?select=temperatures.csv
This data consists of temperatures of INDIA averaging the temperatures of all places month wise. Temperatures values are recorded in CELSIUS

a) Apply Linear Regression using suitable library function and predict the Month-wise temperature.
b) Assess the performance of regression models using MSE, MAE and R-Square metrics
c) Visualize simple regression model.

7. Assignment on Classification technique
Every year many students give the GRE exam to get admission in foreign Universities. The data set contains GRE Scores (out of 340), TOEFL Scores (out of 120), University Rating (out of 5), Statement of Purpose strength (out of 5), Letter of Recommendation strength (out of 5), Undergraduate GPA (out of 10), Research Experience (0=no, 1=yes), Admitted (0=no, 1=yes). Admitted is the target variable.
Data Set: https://www.kaggle.com/mohansacharya/graduate-admissions
The counselor of the firm is supposed check whether the student will get an admission or not based on his/her GRE score and Academic Score. So to help the counselor to take appropriate decisions build a machine learning model classifier using Decision tree to predict whether a student will get admission or not.

a) Apply Data pre-processing (Label Encoding, Data Transformation….) techniques if necessary.

b) Perform data-preparation (Train-Test Split)
c) Apply Machine Learning Algorithm
d) Evaluate Model.

Mini project is to be performed in a group of 3 to 4 students.

Develop a mini project in a group using different predictive models techniques to solve any real life problem.