

**BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI**

**WORK INTEGRATED LEARNING PROGRAMMES**

**COURSE HANDOUT**

**Part A: Content Design**

|  |  |
| --- | --- |
| **Course Title** | Unsupervised Learning and Association Rule Mining |
| **Course No(s)** | PCAM ZC221 |
| **Credit Units** | 2 |
| **Course Author** | S.P. Vimal, Raja Vadhana |
| **Version No** | V1.0 |
| **Date** | 31/08/2018 |

**Course Description**

|  |
| --- |
| Unsupervised learning algorithms find regularities of the given dataset in the absence of explicit labels or supervised outputs for the data points. Clustering is an unsupervised learning task whose objective is to find natural grouping present in the data. This course covers various clustering algorithms like K-Means, EM Algorithm, Single Linkage Algorithm, Complete Linkage algorithm and DBSCAN. Various ways of assessing the quality of clustering and detecting outliers are discussed. The typical industrial applications of unsupervised learning algorithms are covered as well. HMM is introduced in the context of performing time series prediction and the role of EM algorithm in estimating the parameters are discussed. The other part of the course introduces an important class of algorithms to learn association or discover dependencies between the data items, known as learning association rules. We discuss apriori algorithm and different metrics to measure the interestingness of the rules. |

**Course Objectives**

|  |  |
| --- | --- |
| **No** | **Objective** |
| **CO1** | To introduce unsupervised learning, various unsupervised learning algorithms |
| **CO2** | To introduce Association rule learning and apriori algorithms |
| **CO3** | To introduce time series data and use HMM to solve various tasks involving Time series data |

**Text Book(s)**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| T1 | Pattern Recognition and Machine Learning, Christopher M Bishop, 2006, Springer |
| T2 | Data Mining - Concepts and Techniques, Han & Kemper, 3rd Ed, MK |
| T3 | Artificial Intelligence - A Modern Approach, Russel & Norvig, Third Ed, PHI |

**Reference Book(s) & other resources**

|  |  |
| --- | --- |
| No | Author(s), Title, Edition, Publishing House |
| R1 | Introduction to Machine Learning, Ethem Alpaydin, 3rd Ed, 2014, MIT Press |

**Content Structure**

<List down the modular content structure of the course either in the tabular form given below or as bullets>

|  |  |  |
| --- | --- | --- |
| **No** | **Title of the Module** | **Reference** |
| M1 | 1. Introduction to Unsupervised Learning, Clustering    1. Unsupervised Learning - Introduction - Applications- Clustering as an unsupervised learning task - Defining clustering    2. Introducing Various ways to solve clustering problem (similarity based, density based, hierarchical, graph theoretic based) - Notion of quality of clustering    3. Overview of clustering algorithms to be covered in this course 2. Introducing the clustering case study to be used throughout the course for assignments -    1. overview of the data set to be used    2. Exploring data using Python *[To be done by faculty in class]* | Class Notes, T2 |
| M2 | 1. K-Means Algorithm    1. K-Means Algorithm    2. Discussion on Various Initializations, Standardizing Attributes (for eg- z-score) & Convergence    3. Demonstration in Python *[To be done by faculty in class]*    4. Applications of using K-means with Images, videos, documents 2. K-Means - Variations    1. Online stochastic version of k-means (with sequential update) - Discussions on quality of clustering / convergence - Applications    2. Mini-Batch K-Means - Discussions on quality of clustering / convergence - Applications 3. Detecting Outliers    1. Outliers and Clustering - Overview.    2. Using K-means to detect outliers    3. Demonstration in Python *[To be done by faculty in class]* | T1  T1  T2, T1 |
| M3 | 1. EM Algorithm    1. Mixtures of Gaussians (MoG) - Applications, modelled as MoG    2. Using Maximum Likelihood to estimate mixture densities - Issues    3. EM Algorithm for Gaussian mixtures       1. Derivation       2. Illustration (using a problem involving mixture of two gaussians) + Python Demonstration *[To be done by faculty in class]*       3. General Form of EM Algorithm and Applications    4. Relationship to K-Means Algorithm 2. Clustering for Customer Segmentation - [ Industry Talk uploaded over LMS as courseware] | T1, T3  T1, T3 |
| M4 | 1. Hierarchical Clustering    1. Introduction to hierarchical clustering    2. Agglomerative Clustering Vs Divisive Clustering    3. Distance Measures (Minimum distance, Maximum Distance, Mean Distance, Average Distance)    4. Algorithms       1. Single linkage, Complete Linkage algorithm       2. Demonstration in python *[To be done by faculty in class]*       3. Discussion on Termination, efficiency, applications | T2 |
| M5 | 1. Density Based Clustering    1. Density based approach to clustering - Introduction    2. DBSCAN - Density, Density-reachability, Density-connectivity    3. DBSCAN Algorithm    4. Performance & scalability    5. Demonstration using Python *[To be done by faculty in class]* 2. Clustering for Anomaly Detection - *[Industry Talk uploaded over LMS as courseware]* | T2 |
| M6 | 1. Assessing Quality of Clustering    1. Cluster Validity Evaluation (measuring compactness, separation, cluster overlap, etc)    2. Stability of Results from clustering algorithms    3. Determining number of clusters 2. Significance of Clustering - Interpreting/ summarizing Clusters by businesses - *[ Industry Talk uploaded over LMS as courseware]* | T2 |
| M7 | 1. Association Rule Mining    1. Market Basket Analysis - Use cases    2. Terminologies / Measures - association rules, support, confidence ,k-itemset, Frequent itemsets, closed item sets    3. Discussion on computational complexity in generating the itemsets 2. Apriori Algorithm    1. Algorithm    2. Generating Association Rules from frequent itemsets    3. Efficiency Issues and few ways to address it.    4. Evaluating interestingness of patterns    5. Demonstration of Apriori algorithm using python for a practical use case *[To be done by faculty in class]* | R3  R3 |
| M8 | 1. Time series Prediction and Markov Process    1. Introduction       1. Introduction to time series data       2. Time Series prediction applications (eg predicting stock prices, fraud detection, applications in text and speech processing)    2. (discrete) Markov Processes - Overview and Terminologies 2. Hidden Markov Model    * 1. Introduction      2. Evaluation Problem - Given a model, evaluate the probability of observing the sequence - (forward-backward Procedure)      3. Finding most likely state sequence explaining time series data - Viterbi Algorithm      4. Learning Model parameters - An application of EM Algorithm    1. Case Study: Introduce a problem from an application domain- solution using HMM - Python Implementation / Demonstration *[To be done by faculty in class]* | T3  T3 |

**Weekly coverage of the course:**

**Note: The Timing embedded is the effort required from students to complete the respective evaluation components. Below detail only sample assignment designs. Faculty may change exercise/assignment/mini-project design as per the class requirement**

|  |  |
| --- | --- |
| **Week** | **Content / Assignments / Exercises** |
| Week1 | Video Content: M1 (complete), M2.1  Evaluative Quiz: NIL  Exercises (180 mins):   1. Identifying the correct use cases for clustering applications 2. Applying K-means to problems, comment on convergence, various use cases   Assignment (120 mins)::   1. Python program to perform clustering on a given data set using k-means   Minor Projects: Nil |
| Week 2 | Video Content: M2.2, M2.3, M3.1, M3.2(a), M3.2(b)  Evaluative Quiz: Nil  Exercises (120 mins):   1. Sequential update and pkmeans - uses 2. Detect outlier for a sample data set using clustering 3. Model a clustering problem as a MoG, and use ML to solve it.   Assignment : NIL  Minor Projects(180 Min):   1. Use k-means to perform clustering for a given data set 2. Demonstrate the the role of various aspects initialization, learning rate on the quality of optimization 3. Document the findings as a report |
| Week 3 | Video Content: M3.2 (c), M3.2.(d)  Evaluative Quiz: nil  Exercises (150):   1. Exercises on using EM to learn the mixture density for a given MoG   Assignments: (150)   1. Adapt the python program for estimating mixture densities for a different problem   Minor Projects: |
| Week 4 | Video Content: M4, M5.1  Evaluative Quiz: Q1 (preparation time – 300 mins)  Exercises/Assignments: Nil  Assignments: Nil  Minor Projects: Nil |
| Week 5 | Video Content: M5. 2, M6, M7.1  Evaluative Quiz: Nil  Exercises : (~100 Min)   1. Exercises on Identifying interesting rules   Assignments: (( ~200 Min)   1. In Python, evaluate the clustering obtained for a problem (used for the minor project -1) using various measures 2. Write a one page report characterizing quality of clusters, interpreting the clusters   Minor Project : Nil |
| Week 6 | Video Content: M7.2,  Evaluative Quiz: Q2 (preparation time – 300 mins)  Exercises :  Assignment  Minor Projects: Nil |
| Week 7 | Video Content: M.8  Evaluative Quiz: Nil  Exercises/Assignments: Nil  Minor Project 2 (300 mins)   1. For a given time series data, learn HMM, use the HMM for performing prediction in python 2. document how to use the model for predicting the future data, and assess the quality of the results. |

**Evaluation**

**(Changes if any will be announced at the start of the course via LMS. Please check it)**

|  |  |  |
| --- | --- | --- |
| **Evaluation Component** | **Marks** | **Type** |
| Comprehensive Examination | 40% | Closed/Open (TBA) |
| Quizzes (2) | 24% | Open |
| 2 Minor Projects (Evaluated twice) | 24% | Open |
| Assignments/Exercises (2) | 12% | Open |

**Learning Outcomes**

|  |  |
| --- | --- |
| No | Learning Outcomes |
| LO1 | Understand various algorithms for clustering, association rule mining and the role of HMM in time series prediction tasks |
| L02 | Analyze the problem and provide learning solutions using the algorithms covered in this course |
| LO3 | Apply the learning algorithms suitably to solve various tasks including anomaly detection, parameter estimation, segmentation etc. |
| LO4 | Analyze the given problem, decide the suitability of association rule learning technique to solve this and provide a solution |
| LO5 | Apply the HMM suitably to solve problems involving time series data |