**[COC-19009] Data Mining and Machine Learning**

**Teaching Scheme** Lectures: 3 hrs/week

**Examination Scheme** T1, T2 – 20 marks each, End-Sem Exam - 60

**Course Outcomes:**

Students will be able to

1. **Understand various data preprocessing technique**
2. **Select the best features from pattern mining and perform correlation analysis**
3. **Formulate the machine learning problem as classification or prediction**
4. **Study and Compare various classification and prediction algorithms**
5. **Evaluate the machine learning model**

**Unit I Data Mining and Data Preprocessing [10 Hrs]**

**Data mining:** Need of data mining, knowledge discovery from data, kinds of data, data mining techniques, data mining functionalities, issues in data mining

**Input:** Concepts, instances and attributes

**Output:** Knowledge Representation: Decision tables, Decision trees, Decision rules, Rules involving relations, Instance-based representation.

**Data Preprocessing:** Need of data preprocessing, data cleaning, data integration and transformation, data reduction, data discretization and concept hierarchy generation

**Unit II Mining Frequent Patterns, Associations and Correlations [6 Hrs]**

**Basic concepts:** Market Basket Analysis, Frequent Itemsets, Closed Itemsets, and Association Rules, Frequent Pattern Mining

**Apriori algorithm:** Finding frequent itemsets using candidate generation, generating association rules from frequent itemsets

**From Association Mining to Correlation Analysis**: From association analysis to correlation analysis

**Unit III Machine Learning and Statistical Decision Theory [4 Hrs]**

**Machine Learning:** Concept of machine learning, Applications of ML, Design Perspective and Issues in ML, Supervised, Unsupervised, Semi-supervised learning with applications and issues.

**Statistical Decision Theory:** Regression, Classification, Bias Variance

**Classification and Prediction:** Concept, Issues Regarding Classification and Prediction

**Unit IV Classification and Prediction [12 Hrs]**

**Classification by Decision Tree Induction:** Decision Tree representation, Decision Tree Induction, Attribute Selection Measures, Pruning, Rule extraction from Tree, Learning rules from Data

**Bayesian Classification:** Bayes’ Theorem, Naïve Bayesian Classification, Bayesian Belief Networks

**Support Vector Machines:** Objective (optimization), hypothesis, SVM decision boundary, kernel functions in SVM

**Clustering:** Unsupervised learning technique, Similarity and Distance Measures, k-means and k-medoids algorithm, optimization objective, random initialization, choosing value of k

**Prediction:** Linear Regression, Nonlinear Regression

**Unit V Diagnostic and Model Evaluation [6 Hrs]**

**Diagnostic:** Debugging a learning algorithm, evaluating a hypothesis (Model selection), training/validating/testing procedures, diagnosing bias versus variance and vice versa, learning curves, handling the overfitting of the model using regularization techniques

**Accuracy and Error measures:** classifier accuracy measures, predictor error measure, evaluating the accuracy of a classifier or predictor, Confusion metric, precision, recall, tradeoff between both, accuracy, Analysis of ROC

**Text Books:**

1. Tom Mitchell, Machine Learning, McGraw-Hill, 1997

2. Jiawei Han Micheline Kamber, Data Mining Concepts and Techniques, Latest Edition

**References:**

1. Ethem Alpaydin, Introduction to Machine Learning, PHI, 2005
2. D. Hand, H. Mannila and P. Smyth. Principles of Data Mining. Prentice-Hall. 2001
3. K.P. Soman, R. Longonathan and V. Vijay, Machine Learning with SVM and Other Kernel Methods, PHI-2009
4. Christopher M. Bishop, Pattern Recognition and Machine Learning, Springer 2006.
5. M. H. Dunham. Data Mining: Introductory and Advanced Topics. Pearson Education. 2001
6. I. Witten, E. Frank, Mark Hall, C. Pal. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann. 2016
7. T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006. Link: <https://people.inf.elte.hu/kiss/13dwhdm/roc.pdf>

**Online Resources:** Suggested sources for papers are Data Mining and Knowledge Discovery, Transactions on Knowledge and Data Engineering, Knowledge and Information Systems, SIGKDD Explorations, and the KDD, ICDM, ECML/PKDD, PAKDD, and SIAM Data Mining conferences proceedings

**[COC-19012] Data Mining and Machine Learning Laboratory**

**Teaching Scheme: Examination Scheme:**

Practical: 2 hrs/week Term Work: 50 marks

Oral Examination: 50 marks

**Course Outcomes:**

Students will be able to:

1. **Identify the kind of data to be mined and prepare the dataset for analysis**
2. **Compute essential features using appropriate technique**
3. **Apply learning techniques based on classification and prediction to solve a real-life problem.**
4. **Demonstrate the effectiveness of the model using suitable evaluation measures**
5. **Critically analyze the strengths and weaknesses of various Machine Learning approaches**

**Suggested list of Assignments**

**Assignment 1: Dataset selection and Preparation**

Identify and characterize a data set to address the following questions

1. What the data is about?
2. What type of benefit you might hope to get from data mining?
3. What type of data mining (classification, clustering, etc.) you think would be relevant?  
   For each, illustrate with an example, e.g., if you think clustering is relevant, describe what you think a likely cluster might contain and what the real-world meaning would be.
4. Name one type of data mining that you think would *not* be relevant and describe briefly why not.
5. Discuss data quality issues: For each attribute,
   1. Are there problems with the data?
   2. What might be an appropriate response to the quality issues.
6. For at least two attributes, discuss data preprocessing, and give an example of how it would be done / the outcome on a small subset of the data.
   1. What would an appropriate smoothing or generalization technique be?
   2. What is an appropriate normalization or data reduction technique?
7. Apply any 2 Dimension reduction techniques to handle multi-dimensional data.

**Assignment 2: Association Rule Mining**

Select appropriate dataset and Implement *frequent itemset mining* algorithm namely Apriori Algorithm using a programming language that you are familiar with. Compare the performance of algorithm with various kinds of large data sets. Write a report to analyze the situations (such as data size, data distribution, minimal support threshold setting, and pattern density) and give justification.

**Assignment 3: Decision tree classifier**

Implement an algorithm to learn the structure of a binary decision tree. The optimal splits at each node should be found using the information gain criterion. To learn a binary decision tree, you must determine which feature attribute to select as well as the threshold value to use in the split criterion for each non-leaf node in the tree. This can be done in a recursive manner, where we first find the optimal split for the root node using all the training data available to us. We then split the training data according to the criterion selected for the root node, which will leave us with two subsets of the original training data. We then find the optimal split for each of these subsets of data, which gives the criterion for splitting on the second level children nodes. We recursively continue this process until the subsets of training data we are left with at a set of children nodes are pure (i.e., they contain only training examples of one class) or the feature vectors associated with a node are all identical (in which case we can’t split them) but their labels are different. Evaluate the model using appropriate evaluation measures.

**Assignment 4: Bayesian classification**

Implement an algorithm for Bayesian classification and demonstrate its effectiveness on appropriate dataset. Evaluate the model using appropriate evaluation measures. Estimate the accuracy of the classifier using 5-fold cross-validation.

**Assignment 5: Clustering**

Implement any two clustering algorithms. Calculate the accuracy using the handwritten-dataset (<http://yann.lecun.com/exdb/mnist/>). Compare the performance of both algorithms using suitable evaluation measures.

**Assignment 6: Regression**

Choose suitable dataset. Implement linear regression methods namely, Least Squares and Least Mean Squares methods for prediction. Evaluate the model using appropriate evaluation measures.