

B.Sc. III-Year, V-Semester (CBCS) Data Science Syllabus
Paper-V(Theory): Machine Learning
[4 HPW :: 4 Credits :: 100 Marks (External:80, Internal:20)]

Course objectives and Outcomes:

1. Able to know the basic principles and limitations of machine learning.
2. Able to know the various supervised, unsupervised, bio-inspired Probabilistic Modeling and Association Rule Mining.
3. Able to implement the various machine learning algorithms for data analysis using Python and Weka.

UNIT-I

Introduction: Introduction to Machine Learning (ML), Types of ML: Supervised, Unsupervised, Reinforcement, Canonical Learning Problems, The Decision Tree Model of Learning, Formalizing the Learning Problem. **Limits of Learning:** Data Generating Distributions, Inductive Bias, Underfitting and Overfitting, Separation of data to training and test, parameters and Hyperparameters, Real World Applications of Machine Learning. **Geometry and Nearest Neighbors:** From Data to Feature Vectors, K-Nearest Neighbors, Decision Boundaries, K-Means Clustering, High Dimensions.

UNIT-II

Practical Issues: Importance of Good Features, Irrelevant and Redundant Features, Feature Pruning and Normalization, Combinatorial Feature Explosion, Model Evaluation: Cross-Validation, Performance Metrics (Accuracy, Precision, Recall, F1 Score, ROC, AUC), Hypothesis Testing and Statistical Significance, Debugging Learning Algorithms, Bias Variance tradeoff. **Linear Models:** The Optimization Frame work for Linear Models, Convex Surrogate Loss Functions, Weight Regularization, Optimization and Gradient Descent, Support Vector Machines.

UNIT-III

Probabilistic Modeling: Classification by Density Estimation, Statistical Estimation, Naïve Bayes Models, Prediction. **Neural Networks:** Basics of Neural Networks, Activation Functions, The Perceptron Algorithm, Limitations of the Perceptron, Bio-inspired Multi-Layer Networks, The Back-propagation Algorithm, Initialization and Convergence of Neural Networks, Beyond two layers, Breadth vs Depth, Basis Functions.

UNIT-IV

Unsupervised Learning: Clustering Introduction, Similarity and Distance Measures, Agglomerative Algorithms, Divisive Clustering, Minimal Spanning Tree. **Association Rules:** Introduction, large Itemset, Apriori Algorithm.

References:

1. Hal Daume III (2017): A Course in Machine Learning (CIML). (<http://ciml.info/>)
2. Margaret H Dunham (2003): Data Mining: Introductory and Advanced Topics, Pearson Education.
3. Aurélien Géron (2019): Hands on Machine Learning with SciKit-Learn, Keras and Tensor Flow. O'Reily.
4. Chris Albo (2018): Machine Learning with Python Cookbook., O'Reily.
5. Andreas C Miller, Sarah Guido (2017): Introduction to Machine Learning with Python: A guide. O'Reily.

B.Sc. III-Year, V-Semester (CBCS) Data Science Syllabus
Paper-V(Practical): Machine Learning (Lab)
[2 HPW :: 1 Credit :: 25 Marks]

Course objectives and Outcomes:

1. Able to implement the machine learning algorithms: Decision tree, K-means, Perceptron learning, Naïve Bayes, Apriori and agglomerative clustering algorithms using Scikit learn and Weka on the datasets.

List of Practical's:

1. Write a Python program using Scikit-learn to split the Iris dataset consisting of 150 records into 70% train data and 30% test data. Also split data set into the training set will contain 120 records and the test set contains 30 of those records. Print both datasets. Examine the similar and dissimilar records.
2. Write Python program to use sklearn's Decision Tree Classifier to build a decision tree for the sklearn's datasets. Implement functions to find the importance of a split (entropy, information gain, gini measure)
3. Write a Python program to implement your own version of the K-means algorithm. Then apply it to different datasets and evaluate the performance.
4. Design a Perceptron classifier to classify handwritten numerical digits (0-9). Implement using scikit or Weka.
5. Write a Python program to classify text as spam or not spam using the Naïve Bayes Classifier
6. Use WEKA and experiment with the following classifiers: Association Rule Mining (Apriori), Agglomerative and Divisive Clustering.

Note: Students are expected to use machine learning toolkits: Scikit and weka.

1. Scikit-learn(<https://scikit-learn.org/>) an open-source machine learning Python library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.
2. The sklearn datasets package embeds small toy datasets. It includes utilities to load these datasets. It also includes methods to load and fetch popular reference datasets and features some artificial data generators. Students are expected to study and make use of these datasets.
3. Weka (<http://www.cs.waikato.ac.nz/ml/weka/>) is another widely used ML toolkit. Weka also to be implemented on various data sets.

References:

1. Ian Witten, Eibe Frank, and Mark Hall, Chris Pal.: Scikit-learn user guide. https://scikit-learn.org/stable//_downloads/scikit-learn-docs.pdf
2. Morgan Kaufmann: Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition.