|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Sr. No.** | **Units** | **Topics** | **Chapter** | **Reference** | **No. of Lectures** |
| 1. | Introduction | 1.1 - Why Data Mining? 1.2 - What Is Data Mining?, 1.4 - What Kinds of Patterns Can Be Mined? | 1 | 1 | 2 |
| 2. | Data Preprocessing | 2.1- Data Objects and Attribute Types, 2.2 - basic Statistical Descriptions of Data (upto page 51) | 2 | 1 | 2 |
| 3. | 2.2 – Data Quality, 2.3 – Data Preprocessing | 2 | 2 | 6 |
| 4. | Association Analysis | 6.1 – Basic Concepts,6.2- Frequent Itemset Mining Methods (upto page 259) | 6 | 1 | 12 |
| 5. | Classification Analysis | 4.1 – Preliminaries, 4.2 – General Approach to Solving a Classification Problem, 4.3 Decision Tree Induction (Till Pg. 165), 4.5 – Evaluating the Performance of a Classifier | 4 | 2 | 7 |
| 6. | 5.1 – Rule Based Classifier (upto page 212),5.2 – Nearest Neighbor Classifiers, 5.3– Bayesian Classifiers (Complete for discrete data and only introduction of Bayes classifier for continuous attributes) till pg. 233, 5.7.1 – Alternative Metrics | 5 | 2 | 8 |
| 7. | Cluster Analysis | 10.1- Cluster Analysis, 10.2 - Partition Methods, 10.3 - Hierarchical Methods (uptopg 462), 10.4 - Density Based Methods (uptopg 473) | 10 | 1 | 11 |

**B.Sc. (Hons.) Computer Science(w.e.f. 2011)**

**CSHT 616 (IV) Data Mining Guidelines**

**Course Books:**

1. Data Mining: Concepts and Techniques, 3nd edition,Jiawei Han and MichelineKamber
2. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education.

**References:**

1. Data Mining: A Tutorial Based Primer, Richard Roiger, Michael Geatz, Pearson Education 2003.
2. Introduction to Data Mining with Case Studies, G.K. Gupta, PHI 2006
3. Insight into Data mining: Theory and Practice, Soman K. P., DiwakarShyam, Ajay V., PHI 2006

**Practical List:** Practical are to be done using Weka, and a report prepared as per the format\*. The operations are to be performed on built-in dummy data sets of weka and/or the downloadable datasets mentioned in references below. Also wherever applicable, the parameter values are to be varied (upto 3 distinct values). The 'Visualize' tab is to be explored with each operation.

1. Preprocessing : Apply the following filters

weka>filter>supervised>attributed>

* AddClassification
* AttributeSelection
* Discretize
* NominalToBinary

weka>filter>supervised>instance

* StratifiedRemoveFolds
* Resample

weka>filter>unsupervised>attribute>

* Add
* AddExpression
* AddNoise
* Center
* Discretize
* MathExpression
* MergeTwoValues
* NominalToBinary
* NominalToString
* Normalize
* NumericToBinary
* NumericToNominal
* NumericTransform
* PrincipalComponent
* RandomSubset
* Remove
* RemoveType
* ReplaceMissingValues
* Standardize

weka>filter>unsupervised>instance>

* Normalize
* Randomize
* Standardize
* RemoveFrequentValues
* RemoveWithValues
* Resample
* SubsetByExpression

1. Explore the 'select attribute' as follows

weka>attributeSelection>

* FilteredSubsetEval
* WrapperSubsetEval

1. Association mining

weka>associations>

* + - Apriori
    - FPGrowth

1. Classification\*\*

weka>classifiers>bayes>

* NaïveBayes

weka>classifiers>lazy>

* IB1
* IBk

weka>classifiers>trees

* SimpleCart
* RandomTree
* ID3

1. Clustering\*\*

weka>clusters>

* SimpleKMeans
* FarthestFirst algorithm
* DBSCAN
* hierarchicalClusterer

\*Prescribed format:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Dataset | Task | Algorithm | Filter | Parameters | Observations | Inference | Remarks |
| Iris | Preprocessing |  | Unsupervised -> Standardize | ignoreClass: false | For all attributes, mean =0 and standard deviation =1 | The dataset has been standardized |  |
| Iris | Preprocessing |  | Unsupervised -> Normalize | scale: 1.0, translation: 0.0 | The minimum and maximum of all attributes is 0 and 1 respectively | The dataset has been normalized in [0,1] |  |
| Iris | Preprocessing |  | Unsupervised -> Normalize | scale: 2.0, translation: 0.0 | The minimum and maximum of all attributes is 0 and 2 respectively | The dataset has been normalized in [0,2] range |  |

\*\* Proper graphs are to be drawn to compare the accuracies achieved by the variations mentioned below.

* + Applying different algorithms to the same dataset.
  + 10%, 20%, 30%, 40% and 50% Noise.
  + Applying different datasets to the same algorithm.

References for the data sets to be used for the experiments:

1. <http://archive.ics.uci.edu/ml/>
2. <http://www.kdnuggets.com/datasets/index.html>
3. <https://wiki.csc.calpoly.edu/datasets/wiki/apriori> (for Association Mining)