ARSD College, University of Delhi

Model Course Handout/Lesson Plan

Course Name : B.Sc. (Hons.) Computer Science						
Semester	**Course Code**	**Course Title**	**Lecture (L)**	**Tutorial (T)**	**Practical (P)**	**Credit (C)**
VI	32347611	DSE-3 (BSCH17B)-DATA MINING-Practical	4	0	4	6
Teacher/Instructor(s)		Dr. Shalini Gupta				
Session		2021-22				

Course Objective:

- This course introduces data mining techniques and enables students to apply these techniques on real-life datasets.
- The course focuses on three main data mining techniques: Classification, Clustering and Association Rule Mining tasks.

Course Learning Outcomes:

On successful completion of the course, students will be able to do following:
1. Pre-process the data, and perform cleaning and transformation.
2. Apply suitable classification algorithm to train the classifier and evaluate its performance.
3. Apply appropriate clustering algorithm to cluster data and evaluate clustering quality
4. Use association rule mining algorithms and generate frequent item-sets and association rules

Lesson Plan:

<table>
<tr><th>Practical No.</th><th>Practical</th><th>Contact Hrs.</th></tr>
<tr><td>1.</td><td>Create a file “people.txt” with the following data:
<table>
<tr><th>Age</th><th>agegroup</th><th>height</th><th>status</th><th>yearsmarried</th></tr>
<tr><td>21</td><td>adult</td><td>6.0</td><td>single</td><td>-1</td></tr>
<tr><td>2</td><td>child</td><td>3</td><td>married</td><td>0</td></tr>
<tr><td>18</td><td>adult</td><td>5.7</td><td>married</td><td>20</td></tr>
<tr><td>221</td><td>elderly</td><td>5</td><td>widowed</td><td>2</td></tr>
<tr><td>34</td><td>child</td><td>-7</td><td>married</td><td>3</td></tr>
</table>
i) Read the data from the file “people.txt”. ii) Create a ruleset E that contain rules to check for the following conditions: 1. The age should be in the range 0-150. 2. The age should be greater than yearsmarried.</td><td>4</td></tr>
</table>

	3. The status should be married or single or widowed. 4. If age is less than 18 the agegroup should be child, if age is between 18 and 65 the agegroup should be adult, if age is more than 65 the agegroup should be elderly. iii) Check whether ruleset E is violated by the data in the file people.txt. iv) Summarize the results obtained in part (iii) v) Visualize the results obtained in part (iii)	
2.	Perform the following preprocessing tasks on the dirty_iris datasetii. i) Calculate the number and percentage of observations that are complete. ii) Replace all the special values in data with NA. iii) Define these rules in a separate text file and read them. (Use editfile function in R (package editrules). Use similar function in Python). Print the resulting constraint object. – Species should be one of the following values: setosa, versicolor or virginica. – All measured numerical properties of an iris should be positive. – The petal length of an iris is at least 2 times its petal width. – The sepal length of an iris cannot exceed 30 cm. – The sepals of an iris are longer than its petals. iv)Determine how often each rule is broken (violatedEdits). Also summarize and plot the result. v) Find outliers in sepal length using boxplot and boxplot.stats	**4*3=12**
3.	Load the data from wine dataset. Check whether all attributes are standardized or not (mean is 0 and standard deviation is 1). If not, standardize the attributes. Do the same with Iris dataset.	**4**
4.	Run Apriori algorithm to find frequent itemsets and association rules 1.1 Use minimum support as 50% and minimum confidence as 75% 1.2 Use minimum support as 60% and minimum confidence as 60 %	**4*4=16**
5.	Use Naive bayes, K-nearest, and Decision tree classification algorithms and build classifiers. Divide the data set into training and test set. Compare the accuracy of the different classifiers under the following situations: 5.1 a) Training set = 75% Test set = 25% b) Training set = 66.6% (2/3rd of total), Test set = 33.3% 5.2 Training set is chosen by i) hold out method ii) Random subsampling iii) Cross-Validation. Compare the accuracy of the classifiers obtained. 5.3 Data is scaled to standard format.	**4*2=8**
6.	Use Simple Kmeans, DBScan, Hierachical clustering algorithms for clustering. Compare the performance of clusters by changing the parameters involved in the algorithms.	**4*4=16**
	Total	**60**

Evaluation Scheme:

No.	Component	Duration	Marks
1.	Internal Assessment		25
	• Quiz		
	• Class Test		
	• Attendance		
	• Assignment		
2.	End Semester Examination	3 hrs	75

Suggested Books:		
Sl. No.	**Name of Authors/Books/Publishers**	**Year of Publication/Reprint**
1.	Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education.	2006
2.	Data Mining: Concepts and Techniques, 3nd edition,Jiawei Han and Micheline Kamber	2012
3.	Data Mining: A Tutorial Based Primer, Richard Roiger, Michael Geatz, Pearson Education .	2003
4.	Introduction to Data Mining with Case Studies, G.K. Gupta, PHI	2006

Mode of Evaluation:	Internal Assessment / End Semester Exam