1. numericals - register
2. ch2 (from beginning)
3. notepad
4. ch4,5
5. ch1 (1.3,1.4)

B.Sc. (Hons.) Computer Science
Semester VI
BHCS 17B: Data Mining

14-Jan-2022

Sr. No.	**Units**	**Topics**	**Chapter**	**No. of Lectures**
1	Introduction	1.1 - What Is Data Mining? 1.2 Challenges 1.3 Data Mining Origins 1.4 Data Mining Tasks	1	5L
2	Data mining techniques	2.1- Types of data, 2.2 – Data Quality, 2.3.1 Aggregation, 2.3.2 Sampling, 2.3.3 Dimensionality reduction – upto pg 51, 2.3.4 Feature subset selection upto pg 52, 2.4.5 Feature creation upto pg 55, 2.3.6 Discretization upto pg 59, 2.3.7 variable transformations 2.4.3 Dissimilarity among data objects 2.4.4 similarity among data objects	2	10L
3	Classification	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	7L
4		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	8L
5	Association Rules	6.1-Problem definition, 6.2-Frequent itemset generation, 6.3-Rule generation till Pg 351	6	10L
6	Clustering	8.1 Basic concepts of clustering analysis, 8.2 K-Means (8.2.1-8.2.5 except 8.2.3), 8.3 Agglomerative Hierarchical Clustering (except pg 522-524), 8.4 DBSCAN	8	12L

Course Books:

1. Introduction to Data Mining, Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education.

References:

2. Data Mining: Concepts and Techniques, 3nd edition,Jiawei Han and Micheline Kamber

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

5. Insight into Data mining: Theory and Practice, Soman K. P., DiwakarShyam, Ajay V., PHI 2006

Abhi

Nidhi

Anamika

14/1/22

splitting ->
1. binary - 2
2. nominal - binary or multiway
3. ordinal - binary or multiway
4. continuous - binary or multiway(ranges)

1. classifiers -> classification techniques to build classification models
- decision tree, rule-based classifiers, naive bayes classifier
2. performance of a classification model ->
- by using confusion matrix
- by using performance metric {accuracy, error rate}
3. induction tree starts from the node with class label as max value {diff for b,n,o,c}
4. impurity measures -> measures for selecting the best split {diff for b,n and c}
- entropy
- gini
- classification error
5. Gain (when impurity measure = entropy then information gain) -> to determine the goodness of a split.
6. Gain Ratio = Info gain / split info (Entropy is used)

Practical List

The practicals are to be performed on R or Python. The operations are to be performed on downloadable datasets mentioned in references below.

Section 1: Preprocessing

Q1. Create a file “people.txt” with the following data:

Age	agegroup	height	status	yearsmarried
21	adult	6.0	single	-1
2	child	3	married	0
18	adult	5.7	married	20
221	elderly	5	widowed	2
34	child	-7	married	3

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.

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)

Q2. 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

Q3. 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.

Section 2: Data Mining Techniques

Run following algorithms on 2 real datasets and use appropriate evaluation measures to compute correctness of obtained patterns:

Q4. 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 %

Q5. 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.

Q6. Use Simple Kmeans, DBScan, Hierachical clustering algorithms for clustering. Compare the performance of clusters by changing the parameters involved in the algorithms.

Section 3: Project

Q7. Students should be promoted to take up one project on any UCI/kaggle/data.gov.in or a dataset verified by the teacher. Preprocessing steps and at least one data mining technique should be shown on the selected dataset. This will allow the students to have a practical knowledge of how to apply the various skills learnt in the subject for a single problem/project.

Recommended Datasets for Classification[i]:

Abalone, Artificial Characters, Breast Cancer Wisconsin (Diagnostic)

Recommended Datasets for Clustering [ii]:

Grammatical Facial Expressions, HTRU2, Perfume data Recommended Datasets for Association Rule Mining:

The dataset can be downloaded from https://wiki.csc.calpoly.edu/datasets/wiki/apriori (for Association Mining)

[i] http://archive.ics.uci.edu/ml/

ii https://raw.github.com/edwindj/datacleaning/master/data/dirty_iris.csv

Reading material:

1. http://www.dcc.fc.up.pt/~ltorgo/DM1/dataPreProc.html