B. Sc. (H) Computer Science, Semester VI

Data Mining 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` dataset[ii].

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: 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

4.1 Use minimum support as 50% and minimum confidence as 75%

4.2 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.

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

Download 5 datasets from UCI or Kaggle or data.gov.in

Ques1 Data Cleaning

Perform data cleaning: Remove inappropriate data/duplicate data, apply validations,

Ques 2 Preprocessing

1. Remove missing values
2. Remove outliers
3. Apply discretization

Later can be used to find difference in accuracy after applying these data cleaning methods

Ques 3 standardization

Apply different standardization/normalization methods like use of mean or median, use of absolute standard deviation vs normal standard deviation.

Ques 4 Association Rule

Ques 5 Classification

Ques 6 Clustering