DATA MINING LAB

Session-1

1. Create your own EXCEL file. Convert the EXCEL file to .csv format and prepare it as .arff file.

Step 1: Create an Excel file

- Open Microsoft Excel and create a spreadsheet with your desired data.

Step 2: Save as CSV

- Click on "File" in the top left corner.
- Select "Save As."
- Choose the location where you want to save the file.
- In the "Save as type" dropdown menu, select "CSV (Comma delimited) (*.csv)."
- Click "Save."

Step 3: Convert CSV to ARFF using Weka

- Open Weka.
- Click on the "Explorer" tab.
- Under "Preprocess," click on "Open file."
- Select your CSV file.
- Click on the "Open" button.
- In the "Preprocess" panel, click on the "Save" button.
- Choose "Arff" as the file type.
- Save the file with a .arff extension.

4. Preprocess and classify Customer, Argiculture, Weather, Whole-sale Customers or the datasets of your own choice from https://archive.ics.uci.edu/ml/datasets.php

Preprocessing and Classification using Weka:

Step 1: Download a Dataset:

- Visit UCI Machine Learning Repository.
- Choose a dataset relevant to your task (Customer, Agriculture, Weather, Wholesale, etc.).
- Download the dataset.

Step 2: Load Dataset in Weka:

- Open Weka.

- Click on the "Explorer" tab.
- Under "Preprocess," click on "Open file" and select the downloaded dataset.

Step 3: Explore the Dataset:

- Use the "Explorer" panel to understand the attributes and data distribution.
- Handle missing values, if any, using Weka's preprocessing tools.

Step 4: Preprocess Data:

- Address issues like missing values, outliers, or irrelevant attributes.
- Normalize or standardize numeric attributes if needed.
- Encode categorical attributes if required.

Step5: Choose a Classifier:

- Go to the "Classify" tab in Weka.
- Explore different classifiers based on your dataset characteristics (e.g., decision trees, SVM, k-NN).

Session-2

5. Perform the basic pre-processing operations on data relation such as removing an attribute and filter attribute bank data

Step1: Open Weka:

- Launch Weka on your machine.

Step 2: Load Dataset:

- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Select your bank dataset.

Step 3:Explore Attributes:

- Click on the "Attributes" tab to understand the attributes in your dataset.

Step4: Remove an Attribute:

- In the "Preprocess" tab, find the "Remove" filter under the "unsupervised" category.
- Select the "Remove" filter and configure it by specifying the index or name of the attribute you want to remove.
- Click on "Apply" to remove the selected attribute.

Step 5: Filter Attributes:

- You can filter attributes based on different criteria.
- For example, to filter numeric attributes, use the "NumericToNominal" filter under the "unsupervised" category. This converts numeric attributes into nominal ones.
- To filter by range or other criteria, you might use the "Remove" filter or other filters based on your needs.

Step 6: Apply Filters and Save:

- After removing or filtering attributes, click on the "Apply" button in the "Preprocess" tab.
- Optionally, you can save the preprocessed dataset using the "Save" option under the "Preprocess" panel.

6. Demonstrate the preprocessing mechanism on the following datasets:

a. student.arff

b. labor.arff

c. contactlenses.arf

1. **student.arff**

Step 1: Load Dataset:

- Open Weka.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Navigate to the "data" folder within the Weka installation directory.
- Select student.arff and click "Open."

Step 2: Explore Attributes:

- Click on the "Attributes" tab to understand the attributes.

Step 3: Preprocessing Steps:

- Suppose you want to remove the "traveltime" attribute:
  - ✓ In the "Preprocess" tab, find the "Remove" filter under "unsupervised."
  - ✓ Configure the filter by specifying the index or name of the attribute (e.g.,"traveltime").
  - ✓ Click "Apply."

2. labor.arff

Step 1:

- Load Dataset:
- Open Weka.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Navigate to the "data" folder within the Weka installation directory.
- Select labor.arff and click "Open."

Step 2: Explore Attributes:

- Click on the "Attributes" tab to understand the attributes.

Step 3:Preprocessing Steps:

- Suppose you want to normalize numeric attributes:
  - ➢ In the "Preprocess" tab, find the "Normalize" filter under "unsupervised."

- Configure the filter as needed.
- Click "Apply."

3. contactlenses.arff

Step 1:

- Load Dataset:
- Open Weka.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Navigate to the "data" folder within the Weka installation directory.
- Select contactlenses.arff and click "Open."

Step 2: Explore Attributes:

- Click on the "Attributes" tab to understand the attributes.

Step 3:Preprocessing Steps:

- Suppose you want to encode categorical attributes:
  - ✓ In the "Preprocess" tab, find the "NominalToString" or "NominalToBinary" filter under "unsupervised."
  - ✓ Configure the filter as needed.
  - ✓ Click "Apply."

Session-3

8. Implement the Apriori Algorithm to find the association rules in contactless. arff dataset.

Step1: Load the Dataset:

- Open WEKA Explorer.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Load the contactless.arff dataset.

Step 2: Choose Apriori Algorithm:

- Click on the "Associate" tab.
- Under the "Choose" box, select "Apriori."

Step3: Set Parameters:

- Click on the algorithm name (Apriori) to set its parameters.
- Set the minimum support and minimum confidence values based on your requirements.

Step 4:Run Apriori:

- Click "Start" to run the Apriori algorithm.

Step 5: Review Results:

- Explore the generated association rules.
- Analyze the support, confidence, and lift values to understand the strength of the rules.

Session-4

9. Find the frequent patterns using FP-Growth algorithm on contactlenses.arff and test.arff datasets.

Steps 1:Load contactlenses.arff Dataset:

- Open WEKA.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Navigate to the location of contactlenses.arff and load the dataset.

Step 2. Find Frequent Patterns using FP-Growth:

- Click on the "Associate" tab.
- Choose the "FP-Growth" algorithm from the list.
- Set the required parameters, such as "Minimum support."
- Click "Start" to run the FP-Growth algorithm on the dataset.

Step 3. Interpret Results:

- Examine the generated frequent patterns in the "Associations" tab.
- Adjust the minimum support threshold to find patterns with different levels of support.

Step 4. Load test.arff Dataset:

- Follow the same steps as above to load the test.arff dataset.

Step 5. Find Frequent Patterns using FP-Growth for test.arff:

- Repeat steps 2-4, setting the parameters for the test.arff dataset.

10. Generate association rules using Apriori algorithm with Bank.arff relation

a. Set minimum support range as 20% -100% incremental decrease factor as 5% and confidence factor as 80% and generate 5 rules.

b. Set minimum support as 10%, delta 5%, minimum average(4ft) as 150% and generate 4 rules.

a. Set Minimum Support Range:

Step1: Open WEKA.

Step 2: Click on the "Explorer" tab.

Step 3: Under the "Associate" panel, choose the "Apriori" algorithm.

Step 5: Set the minimum support range:

- Start with 20% as the minimum support.
- Incrementally decrease by 5% until reaching 100%.

  Iteration 1: Minimum support = 20%,Iteration 2: Minimum support = 15%,Iteration 3: Minimum support = 10%,Iteration 4: Minimum support = 5%,Iteration 5: Minimum support = 1%
- For each iteration:

  Set the minimum support.

  Set the confidence factor to 80%.

  Click "Start" to run the Apriori algorithm.

  In the "Associations" tab, observe the generated rules and choose the top 5 rules.

b. Set Specific Parameters:

Step 1: Open WEKA.

Step 2: Click on the "Explorer" tab.

Step 3: Under the "Associate" panel, choose the "Apriori" algorithm.

Step 4: Set the following parameters:

- Minimum support: 10%
- Delta: 5%
- Minimum average (4ft): 150%
- Set the confidence factor to 80%.

Step 5: Run the Apriori algorithm:

- Click "Start" to run the Apriori algorithm.
- In the "Associations" tab, observe the generated rules and choose the top 4 rules.

11. Generate association rule for the credit card promotion dataset using a priory algorithm with the support range 40% to 100% confidence as 10% incremental decrease as 5% and generate 6 rules.

Step1: Open WEKA.

Step 2: Click on the "Explorer" tab.

Step 3:Under the "Associate" panel, choose the "Apriori" algorithm.

Step 4:Set Minimum Support Range:

Start with a minimum support of 40%.

Incrementally decrease the minimum support by 5% until reaching 100%.

For each iteration:

Set the minimum support.

Set the confidence factor to 10%.

Click "Start" to run the Apriori algorithm.

In the "Associations" tab, observe the generated rules and choose the top 6 rules.

Session-7

15. Demonstrate the classification rule process on the student.arff, employee. arff and labor.arff datasets using the following algorithms:

a. Logistic Regression
b. Decision Tree
c. Naïve Bayes

Step 1: Load Datasets:

- Open WEKA.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Load the student.arff, employee.arff, and labor.arff datasets.

Step 2. Apply Classification Algorithms:

a. Logistic Regression:

- Click on the "Classify" tab.
- Choose "functions" in the "Choose" box.
- Select "Logistic" as the classifier.
- Click "Start" to apply Logistic Regression.

b. Decision Tree:

- Click on the "Classify" tab.
- Choose "trees" in the "Choose" box.
- Select "J48" as the classifier (Decision Tree).
- Click "Start" to apply Decision Tree.

c. Naïve Bayes:

- Click on the "Classify" tab.
- Choose "bayes" in the "Choose" box.
- Select "NaiveBayes" as the classifier.
- Click "Start" to apply Naïve Bayes.

16. Demonstrate the classification rule process on the student.arff, employee.arff and labor.arff datasets using the following algorithms:

a. K-Nearest Neighbour
b. SVM

Step 1. Load Datasets:

- Open WEKA.
- Click on the "Explorer" tab.
- Under the "Preprocess" panel, click on "Open file."
- Load the student.arff, employee.arff, and labor.arff datasets.

Step 2. Apply Classification Algorithms:

a. K-Nearest Neighbors (KNN):

- Click on the "Classify" tab.
- Choose "lazy" in the "Choose" box.
- Select "IBk" as the classifier (Instance-based learner for KNN).
- Click "Start" to apply KNN.
- Set appropriate options such as the number of neighbors.

b. Support Vector Machine (SVM):

- Click on the "Classify" tab.
- Choose "functions" in the "Choose" box.
- Select "SMO" as the classifier (Support Vector Machine).
- Click "Start" to apply SVM.
- Set appropriate options such as the kernel type, complexity, etc.

Session 8

18. Perform the following:

• Load each dataset into WEKA and perform Naïve-bayes classification and k-Nearest Neighbour classification.

• Interpret the results obtained.

• Plot RoC Curves

• Compare classification results of ID3, J48, Naïve-Bayes and k-NN classifiersfor each dataset, and deduce which classifier is performing best and poor for each dataset and justify

1. Load Datasets and Perform Classification:
   - Open WEKA.
   - Click on the "Explorer" tab.
   - Under the "Preprocess" panel, click on "Open file."
   - Load each dataset for classification.
   - Click on the "Classify" tab.
   - Choose "bayes" for Naïve-Bayes and "lazy" for k-NN in the "Choose" box.
   - Select NaïveBayes and IBk (k-NN) as classifiers.
   - Click "Start" to run the classification.
2. Interpret Results:
   - Switch to the "Result list" tab.
   - Examine the classifier output for Naïve-Bayes and k-NN.
   - Note down important metrics such as accuracy, precision, recall, and F1 score.
3. Plot ROC Curves:
   - Click on the "Classify" tab.
   - Choose "functions" for Naïve-Bayes and "lazy" for k-NN in the "Choose" box.
   - Select NaïveBayes and IBk (k-NN) as classifiers.
   - Click on the "More options..." button.
   - Check the "Output ROC curve" option.
   - Click "Start" to run the classification and plot ROC curves.
4. Compare Classification Results:
   - Compare the classification results of ID3, J48, Naïve-Bayes, and k-NN for each dataset.
   - Consider accuracy, precision, recall, F1 score, and ROC curves.
   - Deduce which classifier is performing best and poor for each dataset.

Session-9

20. Implement simple K-Means Algorithm to demonstrate the clustering rule on the following datasets:

a. iris.arff

b. student.arff

1. Open WEKA Explorer:

- Launch WEKA and open the Explorer.

2. Load Datasets:

a. iris.arff:

- Click on the "Open file" button under the "Preprocess" tab.
- Select the iris.arff dataset.

b. student.arff:

- Similarly, load the student.arff dataset.

3. Choose Cluster Algorithm:

- Click on the "Cluster" tab in WEKA Explorer.
- Under the "Choose" box, select "SimpleKMeans" as the cluster algorithm.

4. Configure K-Means Options:

- Click on the algorithm name "SimpleKMeans" to set its parameters.
- Configure options such as the number of clusters (k) and other settings based on your preferences.
- Distance function as Euclidian
- The number of clusters as 6. With more number of clusters, the sum of squared error will reduce.

5. Run K-Means:

- Click "Start" to run the SimpleKMeans algorithm.

6. View Results:

- After the algorithm completes, go to the "Cluster" tab.
- Explore the results, which may include cluster assignments for instances.

7. Interpretation:

- If visualizations are available, you might observe clusters in graphs or charts.
- Explore the output and understand how instances are grouped into clusters.

21. Perform the following:

• Load each dataset into WEKA and run simple k-means clustering algorithm with different values of k (number of desired clusters).

• Study the clusters formed.

• Observe the sum of squared errors and centroids, and derive insights.

• Explore other clustering techniques available in WEKA.

• Explore visualization features of WEKA to visualize the clusters.

• Derive interesting insights and explain.

1. Load Datasets and Run Simple K-Means:
   - Open WEKA.
   - Click on the "Explorer" tab.
   - Under the "Preprocess" panel, click on "Open file."
   - Load each dataset (e.g., iris.arff and student.arff).
   - Click on the "Cluster" tab.
   - Choose "simple" in the "Choose" box.
   - Select "KMeans" as the clusterer.
2. Run K-Means with Different Values of k:
   - Set different values of k (number of desired clusters).
   - Click "Start" to run the K-Means clustering algorithm for each value of k.
3. Study the Clusters Formed:
   - Switch to the "Cluster mode" tab.
   - Observe the clusters formed for each value of k.
4. Observe Sum of Squared Errors and Centroids:
   - Analyze the sum of squared errors and centroids provided in the output.
   - Understand how these metrics change with different values of k.
5. Explore Other Clustering Techniques in WEKA:
   - Under the "Cluster" tab, explore other clustering algorithms available in WEKA (e.g., hierarchical clustering, DBSCAN, etc.).
   - Run these algorithms on your datasets and compare the results.
6. Explore Visualization Features:
   - Under the "Visualize" tab, explore different visualization options available for clusters.

- Use scatter plots, dendrograms, or other visualizations to observe cluster patterns.

7. Derive Interesting Insights:
   - Analyze the clusters formed with different values of k.
   - Observe how changing k affects the quality of clustering.
   - Compare results across different clustering techniques.

Session-10

22. Implement Hierarchical Clustering Algorithm to demonstrate the clustering rule process in the following datasets;

    a. employee.arff

    b. student.arff

1. Open WEKA:
   - Open the WEKA software.
2. Load Datasets:
   - Click on the "Explorer" tab.
   - Under the "Preprocess" panel, click on "Open file."
   - Load the employee.arff dataset.
   - Repeat the process to load the student.arff dataset.
3. Choose Hierarchical Clustering:
   - Click on the "Cluster" tab.
   - Choose "hierarchical" in the "Choose" box.
   - Select "SimpleKMeans" as the clusterer.
   - Click on the algorithm name to set parameters.
4. Configure Hierarchical Clustering:
   - Set the number of clusters or other relevant parameters.
   - Click "Start" to run the hierarchical clustering algorithm.
5. Explore Results:
   - Switch to the "Cluster mode" tab.
   - Observe the clusters formed.
6. Visualize Dendrogram:
   - Under the "Visualize" tab, choose "Dendrogram" to visualize the hierarchical clustering.
   - Explore the dendrogram to understand the hierarchy of clusters.
7. Derive Insights:

- Analyze the clusters formed by hierarchical clustering.
- Interpret the dendrogram to understand the relationships between data points.

23. Implement Density based Clustering Algorithm to demonstrate the clustering rule process on dataset employee.arff.

Density-based clustering algorithms, like DBSCAN (Density-Based Spatial Clustering of Applications with Noise), are not directly available in WEKA's graphical interface. However, you can use the WEKA Experimenter or command-line interface to run DBSCAN.