

LAB MANUAL

SUBJECT: DATA WAREHOUSING AND MINING
CLASS: T.E (Computer Engineering)
SEMESTER: VI

INDEX

No.	Title	Page No.
1	Introduction to exploratory data analysis using R	3
2	Introduction to regression using R	5
3	Introduction to the Weka machine learning toolkit	7
4	Classification using the Weka toolkit – Part 1	9
5	Classification using the Weka toolkit – Part 2	10
6	Performing data preprocessing for data mining in Weka	12
7	Performing clustering in Weka	15
8	Association rule analysis in Weka	17
9	Data mining case study using the CRISP-DM standard – Part 1	19
10	Data mining case study using the CRISP-DM standard – Part 2	19

Practical one

Title

Introduction to the exploratory data analysis using R

Aim

To learn to perform exploratory data analysis using the R language

References

Venables, Smith, and the R Core Development team, *An Introduction to R*.

Requirements

1. Perform the following tasks
 1. Load the 'iris. CSV' file.
 1. Describe the 'csv' file format briefly.
 2. Describe briefly the characteristics of the iris dataset.
 3. Which command did you use to load the file?
 4. Which command will you use to display the dataset?
 5. Which command will you use to display the names of each column?
 6. Which command will you use to display the names and types of each column?
 2. Descriptive statistics of data
 1. Which command will you use to display the min, max, range, mean, median, variance, standard deviation value of a column of data?
 2. Which command will you use to display the mean value of all numerical columns in the dataset?
 3. Which command will you use to display the mean value of each numerical column grouped by the name of the flower? Write down the results in tabular format and answer the following questions:
 1. Which flower has the lowest mean in petal length?
 2. Which flower has the highest mean in petal length?
 3. Which flower has the lowest mean in petal width?
 4. Which flower has the highest mean in petal width?
 5. Which flower has the lowest mean in sepal length?
 6. Which flower has the lowest mean in sepal width?
 7. Which flower has the highest mean in sepal length?
 8. Which flower has the highest mean in sepal width?
 4. Which command will you use to display the variance of each numerical column grouped by the name of the flower? Write down the results and answer questions as in the previous question.
 5. Use the 'summary' command and write down the types of statistics that are displayed.
 3. Stem-and-Leaf plots and Histograms
 1. Generate stem and leaf plots for sepal length, sepal width, petal length, petal width.
 2. Generate histograms and density plots for each sepal length, sepal width, petal length, petal width?
 1. What do you notice about the distribution of the data for each of these attributes?
 4. Box plots
 1. Generate box plots for each of the numerical attributes
 1. Which attribute has the highest variance?

2. Are there any outliers?
2. Generate boxplots for each of the numerical attributes simultaneously grouped by each flower and answer the following questions?
 1. Which attribute has the highest median?
 2. Which attribute has the highest variance?
5. Quintile quantile plots
 1. Draw quantile-quantile plots for sepal length and petal length.
 2. What is the purpose of drawing quantile-quantile plots?
 3. Which attributes have a normal distribution?
6. Scatter plots
 1. Draw scatter plot of petal length vs. petal width? What can you say about the correlation between these attributes?
 2. Draw scatter plot of petal length vs. petal width? What can you say about the correlation between these attributes?
 3. Draw a scatter plot matrix for the numeric columns.

Practical two

Title

Introduction to linear regression using R

Aim

To learn to perform linear regression using R

References

Venables, Smith, and the R Core Development team, An Introduction to R.

Requirements

Perform the following tasks:

Air Velocity (cm/sec)	20,60,100,140,180,220,260,300,340,380
Evaporation Coefficient (mm²/sec)	0.18, 0.37, 0.35, 0.78, 0.56, 0.75, 1.18, 1.36, 1.17, 1.65

- For the data in the table given above, compute the estimates for the linear regression coefficient estimates manually using the formulas given to you.
- Using R to perform linear regression
 - Draw a scatter plot for the data? Does there appear to be a linear relation?
 - Use R to perform linear regression given the data in section 1.1. Answer the following questions:
 - What command did you use to perform the regression?
 - What command did you use to view the results of the regression?
 - Write the regression formula that was obtained.
 - Is the x-coefficient significant?
 - Is the constant coefficient significant?
 - What the residual standard-error value? What is the significance of this value?
 - What is the R-squared value? What is the significance of this value?
 - Find the correlation coefficient for this data? Which command did you use? What is the significance of the correlation value?
 - What is the significance of the F-statistic?
 - How will you obtain the fitted values for each x-value? Write down the fitted values for each x-value.
 - How will you obtain the residual values for each x-value? Write down the residual values.
 - Use a Quantile-Quantile plot to determine if the residuals are normally distributed? Write down your evaluation of the Quantile-Quantile plot.
- Perform the following tasks:
 - Load the 'baseball.txt' file in R.
 - Perform linear regression on x:bat_ave vs y:homeruns and note down the linear regression equation and other relevant values.
 - Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?
 - Perform a log transformation on the 'homeruns' column, perform linear regression again, and note down the linear regression equation and all relevant values.

5. Create a Quantile-Quantile plot of the residuals? Are the residuals normally distributed?

Practical 3

Title

Introduction to the Weka machine learning toolkit

Aim

To learn to use the Weka machine learning toolkit

References

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

How do you load Weka?

1. What options are available on main panel?
2. What is the purpose of the the following in Weka:
 1. The Explorer
 2. The Knowledge Flow interface
 3. The Experimenter
 4. The command-line interface
3. Describe the arff file format.
4. Press the Explorer button on the main panel and load the weather dataset and answer the following questions
 1. How many instances are there in the dataset?
 2. State the names of the attributes along with their types and values.
 3. What is the class attribute?
 4. In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?
 5. How will you determine how many instances of each class are present in the data
 6. What happens with the Visualize All button is pressed?
 7. How will you view the instances in the dataset? How will you save the changes?
5. What is the purpose of the following in the Explorer Panel?
 1. The Preprocess panel
 1. What are the main sections of the Preprocess panel?
 2. What are the primary sources of data in Weka?
 2. The Classify panel
 3. The Cluster panel
 4. The Associate panel
 5. The Select Attributes panel
 6. The Visualize panel.
6. Load the iris dataset and answer the following questions:
 1. How many instances are there in the dataset?
 2. State the names of the attributes along with their types and values.
 3. What is the class attribute?
 4. In the histogram on the bottom-right, which attributes are plotted on the X,Y-axes? How do you change the attributes plotted on the X,Y-axes?
 5. How will you determine how many instances of each class are present in the data

6. What happens with the Visualize All button is pressed?
7. Load the weather dataset and perform the following tasks:
 1. Use the unsupervised filter RemoveWithValues to remove all instances where the attribute 'humidity' has the value 'high'?
 2. Undo the effect of the filter.
 3. Answer the following questions:
 1. What is meant by filtering in Weka?
 2. Which panel is used for filtering a dataset?
 3. What are the two main types of filters in Weka?
 4. What is the difference between the two types of filters? What is the difference between an attribute filter and an instance filter?
8. Load the iris dataset and perform the following tasks:
 1. Press the Visualize tab to view the Visualizer panel.
 2. What is the purpose of the Visualizer?
 3. Select one panel in the Visualizer and experiment with the buttons on the panel.

Postlab

Provide answers to all the questions given above.

Practical 4

Title

Classification using the Weka toolkit – Part 1

Aim

To perform classification on data sets using the Weka machine learning toolkit

References

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

1. Load the 'weather.nominal.arff' dataset into Weka and run Id3 classification algorithm. Answer the following questions
 1. List the attributes of the given relation along with the type details
 2. Create a table of the weather.nominal.arff data
 3. Study the classifier output and answer the following questions
 1. Draw the decision tree generated by the classifier
 2. Compute the entropy values for each of the attributes
 3. What is the relationship between the attribute entropy values and the nodes of the decision

tree?

4. Draw the confusion matrix? What information does the confusion matrix provide?
5. Describe the Kappa statistic?
6. Describe the following quantities:
 1. TP Rate
 2. FP Rate
 3. Precision
 4. Recall
2. Load the 'weather.arff' dataset in Weka and run the Id3 classification algorithm. What problem do you have and what is the solution?
3. Load the 'weather.arff' dataset in Weka and run the OneR rule generation algorithm. Write the rules that were generated.
4. Load the 'weather.arff' dataset in Weka and run the PRISM rule generation algorithm. Write down the rules that are generated.

Practical 5

Title

Classification using the Weka toolkit – Part 2

Aim

To perform classification on datasets using the Weka toolkit

References

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

1. Load the glass.arff dataset and perform the following tasks?
 1. How many items are there in the dataset?
 2. List the attributes are there in the dataset.
 3. List the classes in the dataset along with the count of instances in the class.
 4. How will you determine the color assigned to each class?
 5. By examining the histogram, how will you determine which attributes should be the most important in classifying the types of glass?
2. Perform the following classification tasks:
 1. Run the 1Bk classifier for various values of K?
 2. What is the accuracy of this classifier for each value of K?
 3. What type of classifier is the 1Bk classifier?
3. Perform the following classification tasks:
 1. Run the J48 classifier
 2. What is the accuracy of this classifier?
 3. What type of classifier is the J48 classifier?
4. Compare the results of the 1Bk and the J48 classifiers. Which is better?
5. Run the J48 and 1Bk classifiers using
 1. the cross-validation strategy with various fold levels. Compare the accuracy results.

2. holdout strategy with three percentage levels. Compare the accuracy results.
6. Perform following tasks:
 1. Remove instances belonging to the following classes:
 1. build wind float
 2. build wind non-float
 2. Perform classification using the 1Bk and J48 classifiers. What is the effect of this filter on the accuracy of the classifiers?
7. Perform the following tasks:
 1. Run the J48 and the NaiveBayes classifiers on the following datasets and determine the accuracy:
 1. vehicle.arff
 2. kr-vs-kp.arff
 3. glass.arff
 4. wave-form-5000.arffOn which datasets does the NaiveBayes perform better? Why?
8. Perform the following tasks
 1. Use the results of the J48 classifier to determine the most important attributes
 2. Remove the least important attributes
 3. Run the J48 and 1Bk classifiers and determine the effect of this change on the accuracy of these classifiers. What will you conclude from the results?

Practical 6

Title

Performing data preprocessing tasks for data mining in Weka

Aim

To learn how to use various data preprocessing methods as a part of the data mining

References

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

Part A: Application of Discretization Filters

Perform the following tasks

1. Load the 'sick.arff' dataset
2. How many instances does this dataset have?
3. How many attributes does it have?
4. Which is the class attribute and what are the characteristics of this attribute?
5. How many attributes are numeric? What are the attribute indexes of the numerical attributes?
6. Apply the Naive Bayes classifier. What is the accuracy of the classifier?
2. Perform the following tasks:
 1. Load the 'sick.arff' dataset.
 2. Apply the supervised discretization filter.
 3. What is the effect of this filter on the attributes?
 4. How many distinct ranges have been created for each attribute?
 5. Undo the filter applied in the previous step.
 6. Apply the unsupervised discretization filter. Do this twice:
 1. In this step, set 'bins'=5
 2. In this step, set 'bins'=10
 3. What is the effect of the unsupervised filter on the dataset?
 7. Run the Naive Bayes classifier after applying the following filters
 1. Unsupervised discretized with 'bins'=5
 2. Unsupervised discretized with 'bins'=10
 3. Unsupervised discretized with 'bins'=20.
 8. Compare the accuracy of the following cases
 1. Naive Bayes without discretization filters
 2. Naive Bayes with a supervised discretization filter
 3. Naive Bayes with an unsupervised discretization filter with different values for the 'bins' attributes.

Part B: Attribute Selection

1. Perform the following tasks:
 1. Load the 'mushroom.arff' dataset
 2. Run the J48, 1Bk, and the Naive Bayes classifiers.
 3. What is the accuracy of each of these classifiers?
2. Perform the following tasks:

1. Go to the 'Select Attributes' panel
2. Set attribute evaluator to CFSSubsetEval
3. Set the search method to 'Greedy Stepwise'
4. Analyze the results window
5. Record the attribute numbers of the most important attributes
6. Run the meta classifier AttributeSelectedClassifier using the following:
 1. CFSSubsetEval
 2. GreedStepwise
 3. J48, 1Bk, and NaiveBayes
7. Record the accuracy of the classifiers
8. What are the benefits of attribute selection?

Part C

1. Perform the following tasks:
 1. Load the 'vote.arff' dataset.
 2. Run the J48, 1Bk, and Naive Bayes classifiers.
 3. Record the accuracies.
2. Perform the following tasks:
 1. Go to the 'Select Attributes' panel
 2. Set attribute evaluator to 'WrapperSubsetEval'
 3. Set search method to "RankSearch"
 4. Set attribute evaluator to 'InfoGainAttributeEval'
 5. Analyze the results
 6. Run the meta classifier AttributeSelectedClassifier using the following:
 1. WrapperSubsetEval
 2. RankSearch
 3. InfoGainAttributeEval
 7. Sampling
 1. Load the 'letter.arff' dataset
 2. Take any attribute and record the min, max, mean, and standard deviation of the attribute
 3. Apply the Resample filter with 'sampleSizePercent' set to 50 percent
 4. What is the size of the filtered dataset. Observe the min, max, mean, and standard deviation of the attribute that was selected in step 2. What is the percentage change in the values?
 5. Give the benefit of sampling a large dataset.

Practical 7

Title

Performing clustering using the data mining toolkit

Aim

To learn to use clustering techniques

References

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

Part 1

Perform the following tasks:

1. Load the 'bank.arff' data set in Weka.
2. Write down the following details regarding the attributes:
 1. names
 2. types
 3. values.
3. Run the SimpleKMeans clustering algorithm on the dataset
 1. How many clusters are created?
 2. What are the number of instances and percentage figures in each cluster?
 3. What is the number of iterations that were required?
 4. What is the sum of squared errors? What does it represent ?
 5. Tabulate the characteristics of the centroid of each cluster.
 6. Visualize the results of this clustering (let the X-axis represent the cluster name, and the Y-axis represent the instance number)
 1. Is there a significant variation in age between clusters?
 2. Which clusters are predominated by males and which clusters are predominated by females?
 3. What can be said about the values of the region attribute in each cluster?
 4. What can be said about the variation of income between clusters?
 5. Which clusters are dominated by married people and which clusters are dominated by unmarried people?
 6. How do the clusters differ with respect to the number of children?
 7. Which cluster has the highest number of people with cars?
 8. Which clusters are predominated by people with savings accounts?
 9. What can be said about the variation of current accounts between clusters?
 10. What can be said about the variation of mortgage holdings between clusters?
 11. Which clusters comprise mostly of people who buy the PEP product and which ones are comprised of people who do not buy the PEP product?
4. Run the SimpleKMeans algorithm for values of K (no. of clusters) ranging from 1 to 12. Tabulate the sum of squared errors for each run. What do you observe about the trend of the sum of squared errors?
5. For the run with K=12, answer the following questions:
 1. Is there a significant variation in age between clusters?
 2. Which clusters are predominated by males and which clusters are predominated by females?

3. How do the clusters differ with respect to the number of children?
4. Which clusters comprise of people who buy the PEP product and which ones are comprised of people who do not buy the PEP product?
5. Do you see any differences in your ability to evaluate the characteristics of clusters generated for $K=6$ versus $K=12$? Why does this difference arise?

Part 2

1. Perform the following tasks:
 1. Load the 'iris.arff' dataset
 2. Write down the following:
 1. The names of the attributes
 2. The types of the attributes
 3. The class attribute, its type, and possible values.
 3. Run the SimpleKMeans clustering algorithm on the dataset as follows
 1. Set $K=2$ and observe the sum of squared errors.
 2. Set $K=3$ and observe the sum of squared errors.
 3. Set $K=4$ and observe the sum of squared errors.
 4. Set $K=5$ and observe the sum of squared errors.
 5. What can be said about the trend of the sum of squared errors? What does this trend imply
 6. For $K=3$, tabulate the characteristics of each centroid? How do the clusters correspond to the class values?

Practical 8

Title

Using Weka to determine Association rules

Aim

To learn to use Association algorithms on datasets

References

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

1. Perform the following tasks
 1. Define the following terms
 1. item and itemset
 2. Association
 3. Association rule
 4. Support of an association rule
 5. Confidence of an association rule

6. Large itemset
7. Association rule problem
2. What is the purpose of the Apriori algorithm
2. Perform the following tasks:
 1. Load the 'vote.arff' dataset
 2. Apply the Apriori association rule
 3. What is the support threshold used? What is the confidence threshold used?
 4. Write down the top 6 rules along with the support and confidence values.
 5. What does the figure to the left of the arrow in the association rule represent?
 6. What does the figure to the right of the arrow in the association rule represent?
 7. For rule 8, verify that numerical values used for computation of support and confidence are in accordance with the data by using the Preprocess panel. Then compute the support and confidence values. Are they above the threshold values?
3. Perform the following tasks:
 1. Load the dataset 'weather.nominal.arff'.
 2. Apply the Apriori association rule
 1. Consider the rule "temperature=hot ==> humidity=normal." Compute the support and confidence for this rule.
 2. Consider the rule "temperature=hot humidity=high ==> windy=TRUE." Consider the support and confidence for this rule.
 3. Is it possible to have a rule like the following rule:
"outlook=sunny temperature=cool" ==> humidity=normal play=yes
4. Perform the following tasks:
 1. Load the bank-data.csv file.
 2. Apply the Apriori association rule algorithm. What is the result? Why?
 3. Apply the supervised discretization filter to the age and income attributes.
 4. Run the Apriori rule algorithm
 5. List the rules that were generated.

Practical 9-10

Title

Part 1: Data mining case study using the CRISP-DM standard: business understanding, data understanding, data preparation.

Part 2: Data mining case study using the CRISP-DM standard: modeling, evaluation, deployment.

Aim

To perform the following phases of the CRISP-DM standard:

1. Business understanding
2. Data understanding
3. Data preparation.

References

Larose, Daniel, *Data Mining: Methods and Model*, John Wiley & Sons. 2006.

Witten, Ian and Eibe, Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Springer.

Requirements

1. CRISP-DM
 1. Briefly describe the CRISP-DM standard.
2. Business understanding phase:
 1. Enunciate project objectives and requirements
 2. Translate project objectives into a data mining task
 1. Specify the data mining task
 2. Provide a cost-benefit table
 3. Prepare a preliminary strategy.
3. Data understanding phase:
 1. Collect the data
 1. Specify the characteristics of the ClothingStore dataset
 2. Use exploratory data analysis to familiarize yourself with the data and discover initial insights
 1. ClustType attribute analysis
 1. Is there any pattern relating cluster type to the response variable?
 2. What are the six largest clusters?
 2. CustomerId attribute analysis
 1. Should this be an important field in the analysis?
 3. ZipCode attribute analysis
 1. Should this be an important field in the analysis?
 4. Evaluate the following fields
 1. Rec (number of purchase visits)
 2. Fre (frequency of visits)
 3. Mon (Total spending)
 4. Avg (average spent per visit)
 5. Days (Days on file)
 6. Fredays (days between purchase visits)
 7. Classes (number of different classes purchased)
 8. Styles (number different styles purchased)

9. Stores (number of stores visited)
 10. StoLoy (store loyalty)
 11. Hi (product uniformity)
 12. Ltfreday (lifetime average time between visits)
- Do you find any relationship between these fields and the response?
 What is the type of distribution for the attributes?
 Apply the log transform to these fields.
5. Evaluate the following fields:
 1. Psweaters
 2. Pknittops
 3. and all other similar fields

Do you find any relationship between these attributes and the target variable?
 What type of distributions do you have here?
 6. Standardize all the numeric fields.
 7. Evaluate the following fields to identify the relationship between predictors and response variable
 1. Rec (Number of purchase visits)
 2. Fre (frequency of visits)
 3. Mon (total net sales)
 4. Avrg (average spent per visit)
 5. Classes (number of different product classes purchased)
 6. Coupons (number of coupons used)
 7. Hi (product uniformity)
 8. Ltfredays (lifetime average time between visits)

Write down your conclusions.
 8. Evaluate the quality of the data
3. If desired, select interesting subsets that may contain actionable patterns.
4. Data preparation phase
 1. Select the cases and variables you want to analyze and that are appropriate for your analysis
 1. Write down the fields that you feel are very important in the analysis
 2. Perform transformation on variables, if needed
 1. Normalization of the variables has already been performed
 3. Clean raw data so that it is ready for the modeling tools.
 1. The initial data provided is clean
 5. Modeling Phase
 1. Select and apply appropriate modeling techniques
 1. Clustering
 1. Apply SimpleKMeans clustering to obtain 2 clusters? Which cluster has the highest response rate? What is the percentage of records that belong to each cluster?
 2. Apply SimpleKMeans clustering to obtain 3 clusters? Which cluster has the highest response rate? What is the percentage of records that belong to each cluster?
 3. Remove all fields related to Response? Apply SimpleKMeans clustering to obtain 2 clusters? What is the percentage of records in each cluster?
 4. Visualize the clusters and write down the relevant observations.
 2. Classification
 1. Based on your observations, eliminate attributes that are not important in distinguish between the response attributes.
 2. Run the J48 decision tree algorithm

1. Which attributes are selected to be most important? Write down the classification accuracy results
3. Run the logistic regression model. Write down the equation that is obtained. Write down the classification accuracy results.
4. Run the neural network model. Write down relevant details regarding the neural network that was created.
2. Calibrate model settings to optimize results
 1. Not required
 2. Often, several different techniques may be applied for the same data mining problem
 1. Not required
 3. Loop back to the data preparation phase as required
 1. Not required
6. Evaluation Phase
 1. Evaluate the the models developed.
 1. Present the accuracy results for each of the classifiers
 2. Determine if the model achieves effectiveness
 1. Not required
 3. Establish whether some facet of the business or search problem has not been accounted for sufficiently.
 4. Finally, come to a decision regarding the use of data mining results
 1. Provide your conclusions
7. Deployment Phase
 1. Model creation does not signify the completion of the project. Need to make use of created models according to business objectives.
 1. Not required.
 2. Example of simple deployment: Generate a report
 1. Not required.
 3. Example of a more complex deployment: Implement a parallel data mining process in another department.
 1. Not required.
 4. For businesses, the customer often carries out the deployment based on your model.
 1. Not required.

Sources