Department of Computer Science, Banaras Hindu University, Varanasi – 221005.

Paper No.: CS311: Data Mining Practical

Data mining laboratory assignment questions

Note: Any standard dataset (UCI repository or from any other sources) can be used for experiments. Students also free to use any Tool/programming language/software. Python/WEKA tool is preferred and the documentation can be found at: http://www.cs.waikato.ac.nz/ml/weka/

1. Data preprocessing and Visualization for data mining: select a simple numerical dataset and do the following.
   i. Apply data cleaning techniques (missing values, binning, smoothing, outlier etc.).
   ii. Find the mean, median, mode, max, min, quantiles, outliers, standard deviation and variance of the data.
   iii. Find the first quartile (Q1), the third quartile (Q3) of the data and Inter-quartile range.
   iv. Give the five-number summary of the data.
   v. Show a histogram, boxplot, quantile plot, quantile-quantile (Q-Q) plot and scatter plot of the data.

2. Implementation of *Apriori* algorithm for finding *Frequent Itemsets* and *Association Rules*.
   i. Select a suitable dataset and do the required preprocessing.
   ii. Find the all frequent itemsets using *Apriori* algorithm for various *support* and *confidence*.
   iii. Generate *Association Rules* from *Frequent Itemsets.*

3. Implementation of *FP-growth* algorithm for finding *Frequent Itemsets* and *Association Rules*.
   i. Select a suitable dataset and do the required preprocessing.
   ii. Find the all frequent itemsets using *FP-growth* algorithm for various *support* and *confidence*.
   iii. Generate *Association Rules* from *Frequent Itemsets.*

4. Construct the decision tree using CART algorithm and evaluate the performance.
   i. Select a suitable data set and do the required preprocessing.
   ii. Construct the tree using CART algorithm.
   iii. Perform cross validation and tune the parameters to improve model's overall performance.
   iv. Use different attribute selection measures and compare the model performance.
   v. Visualize the constructed decision tree.
   vi. Extract rules form the constructed tree

5. Construct the decision tree using C4.5 algorithm and evaluate the performance.
   i. Select a suitable data set and do the required preprocessing.
   ii. Construct the tree using C4.5 algorithm.
   iii. Perform cross validation and tune the parameters to improve model's overall performance.
   iv. Use different attribute selection measures and compare the model performance.
   v. Visualize the constructed decision tree.
   vi. Extract rules form the constructed tree

6. Construct the naïve bayes classifier and KNN classifiers and classify the given data sample.
   i. Select a suitable data set and do the required preprocessing.
   ii. Construct the naïve bayes classifier and evaluate its performance using cross validation.
   iii. Find out the optimal *k* value of KNN algorithm for the dataset.
   iv. Classify the given unknown data sample using the constructed models.

7. Implementation of K-Means and K-Medoids.

i. Select a suitable data set and do the required preprocessing.
ii. Apply K-Means clustering algorithm for different values of *k* and present the results.
iii. Apply K-Medoids clustering algorithm for different values of *k* and present the results.
iv. Add some noise/outlier in the dataset and compare its effects on the results of both methods.

8. Implementation of EM and Density Based clustering methods.
   i. Select a suitable data set and do the required preprocessing.
   ii. Cluster the dataset with Expectation Maximization (EM) algorithm and present the results.
   iii. Apply Density Based clustering method for different parameter values and present the results.
   iv. Add some noise/outlier in the dataset and compare its effects on the results of both methods.

9. Implementation of linear regression.
   i. Select a suitable data set and do the required preprocessing.
   ii. Apply linear regression and write down the learned regression function.
   iii. Evaluate the model with the following metrics using cross validation:
      a) R Square/Adjusted R Square.
      b) Mean Square Error(MSE)/Root Mean Square Error(RMSE)
      c) Mean Absolute Error(MAE)
   iv. Visualize the errors made by the learned regression function.

10. Text mining: Build classifiers for the two training sets using (1) SVM and (2) NaiveBayesMultinomial, evaluating them on the corresponding test set in each case.
    i. Select a suitable data set and do the required preprocessing.
    ii. Evaluate the models and interpret the results.
    iii. Compare the results for classification with and without attribute selection.
    iv. Classify unknown instances.
