IITM-CS5011 : Introduction to Machine Learning — Given on: Oct 20, 5pm

Programming Assignment #3 — Due on : Nov 16 , 10pm

- The goal of this assignment is to experiment with decision trees for classification and several clustering techniques.
- This is an individual assignment. Collaborations and discussions with others are strictly prohibited.
- You need to use Weka for this assignment.
- You have to turn in a detailed report of the results of the experiment electronically in Moodle. Typeset your report in Latex.
- Be precise for your explanations in the report. Unnecessary verbosity will be penalized.
- You have to check the Moodle discussion forum regularly for updates regarding the assignment.

1. You have been provided with the following 8 2-dimensional datasets for clustering: Aggregation, compound, path-based, spiral, D31, R15,Jain, Flames. First two columns are the features and the third column is the class label. In all your experiments, make sure that you are not giving the third column also as input to the clustering algorithm. You need to turn in the visualizations of your results for each question.
   (a) Convert all 8 datasets into ARFF format.
   (b) Visualize all 8 datasets. You need to turn in all your plots. Analyze each dataset by visualization and explain how these clustering algorithms will perform in these data (with reasons) : K-means, DBSCAN, hierarchical clustering with single link and complete link.
   (c) Run K-means with R15 dataset. Set k = 8. Report the cluster purity. Vary the value of k from 1 to 20 and study the effect of k on cluster purity. Plot a graph which explains your study.
   (d) Run DBSCAN with Jain dataset. Again report cluster purity. Study the effect of minpoints and epsilon on cluster purity.
   (e) Run DBSCAN and hierarchical clustering with path-based, spiral and flames. Compare their performance in each dataset. For hierarchical you need to experiment with all types of linkages available in weka to find the one that best suits the data.
   (f) Run K-means with D31 dataset. Can you recover all 31 clusters with k=32? If not, can you recover all clusters by increasing the value of k? What happens when you apply DBSCAN? Apply hierarchical clustering with Ward's linkage. How does it perform?

Submission Instructions

Submit a single tarball/zip file containing the following files in the specified directory structure. Use the following naming convention: 'cs5011_a3_rollno.tar.gz'.

cs5011_a3_rollno

Dataset
  spiral.arff
  ...

Report
  rollno-report.pdf

Code
  all your code files