CS5011 : Introduction to Machine Learning

Programming Assignment #3

- The goal of this assignment is to understand the concepts of decision trees and clustering techniques.
- This is an individual assignment. Collaborations and discussions with others are strictly prohibited.
- You need to use Weka for this assignment.(No Exceptions for this)
- You have to turn in the well documented code along with a detailed report of the results of the experiment electronically in Moodle. Typeset your report in LaTeX.
- Your report should contain detailed answers for all of the questions asked below.
- Look at the end of the assignment for submission instructions.
- Submission deadline: 6th November, 2015.
- **No late submissions** will be entertained for this assignment.

Clustering

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.

1. Convert all 8 datasets into ARFF format.
2. 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 on these datasets (with reasons) : K-means, DBSCAN, hierarchical clustering with single link and complete link.
3. Run K-means on 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.
4. Run DBSCAN on Jain dataset. Again report cluster purity. Study the effect of *minpoints* and *epsilon* on cluster purity.

5. Run DBSCAN and hierarchical clustering on Path-based, Spiral and Flames datasets. Compare their performance on each dataset. For hierarchical clustering, you need to experiment with all types of linkages available in Weka to find the one that best suits the data.

6. Run K-means on 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 run DBSCAN on the dataset? Run hierarchical clustering with Wards linkage on the dataset. How does it perform?

Decision Trees

For this experiment, we will use Mushroom dataset from UCI machine learning repository (`https://archive.ics.uci.edu/ml/datasets/Mushroom`). This is a 2-class problem with 8124 instances. Use the last 1124 instances as test data and the rest as training data.

1. Convert the data into ARFF format.

2. Run J48 Decision Tree algorithm from Weka. Report precision, recall and f1- measure. What is the effect of $MinNumObj$ on the performance? What happens when you do $reducedErrorPruning$?

3. What are the important features in deciding whether a mushroom is edible or not?

4. Turn in the Decision Tree learnt by the model (the decision tree with the best performance).

Submission Instructions

Submit a single tarball file containing the following files in the specified directory structure. Use the following naming convention: 'cs5011_pa3_rollno.tar.gz'

data
    spiral.arff
    ...
code
    all your code files
report.pdf
decision tree model