Assignment 9: Classification and Clustering in R

CS 69001: Computing Lab 1

Introduction:

In this assignment you will learn to use classification and clustering methods in R.

Objective:

The aim of this assignment is to get acquainted with R functionality for classification and clustering.

Problem Description:

Classification and clustering are two of the most commonly used data mining techniques. In classification, a classifier is initially trained using a set of labeled data in a supervised manner. Once the classifier is trained, it can be used to determine the class of an unknown sample. On the other hand, clustering is used to determine natural groupings in a given data set in an unsupervised manner.

One of the most common algorithms for clustering is k-means clustering. [Refer to tutorials for more details]. Given a data set and a value of k, the goal is to group the data elements into k clusters so that a particular measure of clustering error (squared sum of distance from cluster centers over all data points) is minimized.

A well-known classification technique is to use Support Vector Machines (SVMs). [Refer to tutorials for more details]. SVMs can be implemented using various types of kernels like linear, quadratic, radial basis function (RBF). Depending on the application and the nature of the data set, classification performance can be influenced by the choice of kernel.

Data sets:
University of California, Irvine has a machine learning repository (https://archive.ics.uci.edu/ml/datasets.html), which is an extensive collection of rich data sets that are quite commonly used to test the performance of machine learning algorithms [Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.]

In this assignment, we will use two of the data sets:

For classification: Bank Marketing Data Set
https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
[Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

For clustering: Individual household electric power consumption Data Set (https://archive.ics.uci.edu/ml/datasets/Individual+household+electric+power+consumption#)

Go through the data set descriptions to get an idea about their content, data format, etc.

Task 1: For classification, consider all or a subset of the features you feel (or experimentally determine) pertinent for classification. Use R to develop the code for SVM based classification using Linear and RBF kernels. The method-feature set combination that gives the best classification accuracy is the one you need to submit, i.e., in your code you need to specify which subset of features you are using and which kernel you are employing for the SVM. You need not submit the other codes. Whoever comes up with the highest accuracy using the least number of features will get some additional bonus marks. You must use 70% of the data for training and the remaining 30% for testing. No data can be ignored. **[50 marks]**

Task 2: For clustering, ignore the date and time features in the data. Also, identify and ignore the entries with missing values. Use R to develop the code for k means clustering. First, consider all the remaining features for clustering. Then, consider the last three features (7-9) only. In each case, vary the value of k starting from 5 to 20. Note clustering accuracy and execution time. Generate input file to display variation of accuracy and execution time with the value of k in Graphviz (for the two cases – all features and only the last three features). Submit your R code which should take the value of k and the set of considered features (Using the attribute numbers mentioned in the data set) as input. It is OK to hardcode the data set name and parsing of the data in your program. For example, if we want to run your code using feature numbers 3-9 with k= 6, we will pass 6 and 3,4,5,6,7,8,9 as parameters. If we want to run your code using feature numbers 7-9 with k=10, we will pass 10 and 7,8,9 as parameters.
[50 Marks]

Deliverables:

You will submit your R programs in a single rollno_a9.tar.gz file in the Moodle submission link. It is mandatory that your code should follow proper indentation and commenting style. There will be deductions in the awarded marks, if you fail to do so.