COMP47460 Assignment 3

Deadline: Submit no later than **20th Dec, 2020**.

Instructions

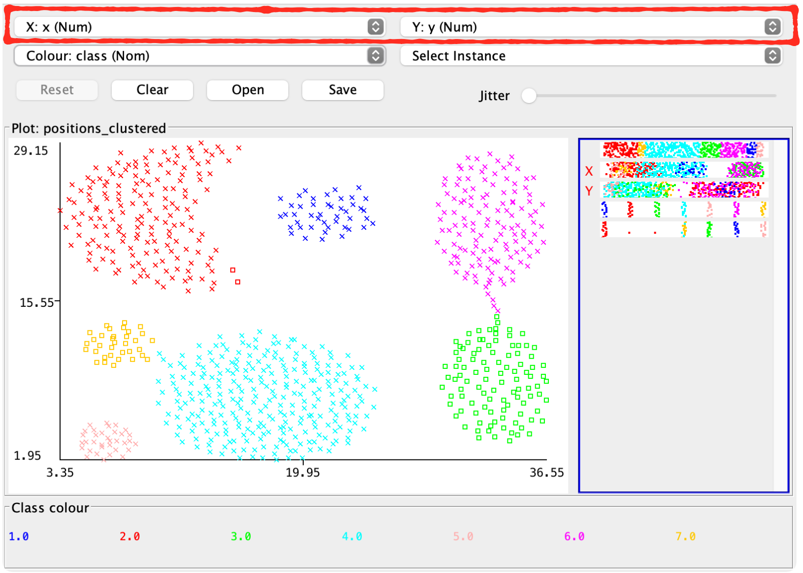
Answer both questions. Submit your assignment as one PDF file (not a DOC/DOCX/ODT/ZIP file) via the module Brightspace page. Include your full name and student ID number on the PDF.

This assignment should be completed individually. Any evidence of plagiarism will be reported to the CS plagiarism committee and it can result in a Fail grade.

Please stick to the suggested page lengths.

Question 1

This questions investigates some aspects of clustering. The dataset you are provided with has 7 clusters which are shown below:



X: x (Num)

Y: y (Num)

Colour: class (Nom)

Select Instance

Plot: positions_clustered

Make sure you select the *x* and *y* numeric attributes to visualise the cluster.

The "Colour: class (Nom)" is the ground truth cluster label (shown here).

After applying a clustering algorithm you will see a "Colour: Cluster (Nom)", which is the label found by the clustering algorithm.

Download the file cluster_a3.arff from Brightspace (there is only one data file).

This is an open-ended assignment -- You do not have to restrict yourself to what is asked. You can take the exploration and the discussion deeper than what is asked.

Total suggested page length for Q1 is 4-5 pages.

(a) In the 'Cluster' tab in Weka, apply the 'SimpleKMeans' clusterer. In the 'Cluster Mode' window, choose the "Classes to clusters evaluation" method. Run the "SimpleKMeans" clusterer with the default parameter settings.

(a) Explain how Weka evaluates the quality of the clustering. [10 marks]

(b) Determine the cluster quality for values of k from 2 to 10. [10 marks]

(c) Discuss the quality of the clustering when k=7 and you change the initialisation method. Use each of the initialisation methods: 'Random', 'k-means++', 'Farthest first'. [10 marks]

(b) Now switch the clusterer to "HierarchicalClusterer" and set k=7. Compare the resulting clusters produced with these linkages: "Single", "Complete", "Average", "Centroid". Specifically:

(a) You should discuss the quality of the clustering. [5marks]

(b) Make some qualitative assessments of the differences in the clusters produced by different linkage types (using the 'Visualise cluster assignments' in Weka. [5marks]

(c) Comment on the difference between the quality of the clusters with k=7 produced by "SimpleKMeans" and "HierarchicalClusterer". [10marks]

(d) Consider the trees produced each linkage type using 'Visualise tree' in Weka and comment on how the tree structure relates to the cluster assignment. [10 marks]

Question 2

Answers all parts below. **Please provide answers in your own words.**

Total page length for Q2 should not exceed 4 pages

(a) You are required to apply dimension reductions techniques on your dataset (eg. PCA). How do you evaluate the quality of the dimension reduction method? [8 marks]

(b) After spending several hours, you are anxious to build a high accuracy model. You built5 boosting models, but neither of these models performed better than benchmark score.Finally, you decided to combine those models as ensemble models are known to provide high accuracy. If your accuracy still doesn't improve, what could potentially be wrong with your ensemble model? [8 marks]

(c) Discuss some of the reasons why you might need a validation set in addition to the training and test set. [8 marks]

(d) When building a classification model you notice your dataset has high variance. How does this affect your choice of model? [8 marks]

(e) Explain the difference between batch gradient descent and stochastic gradient descent. [8 marks]

Grading Guideline

- Q1: 60 marks
- Q2: 40 marks