University of Canberra
Faculty of Science and Technology

Programming for Data Science 11521G (Online & On-campus)

Assignment 1 Marking Guideline
Classifier and Cluster Analysis in Data Science

Due dates: 23:59 Sunday 19/09/2020 (Week 7)
Mark for assessment: 100 points (20% of final mark)
Submission: Submit a .zip file containing all Python files (.py) in your project via Canvas site.
Late submission: 5% of the total mark per day (10 marks per day). Information on how to apply for extension can be found in the unit outline on Canvas.

Question 1 [30 marks]

- [2.5 marks] Read 3 files for red, green, and unknown data sets
- For each unknown sample in the unknown data set
    - [2.5 marks] Calculate distances from the unknown sample to all red data samples
    - [2.5 marks] Find min_1 (minimum distance of the above distances to red samples)
    - [2.5 marks] Calculate distances from the unknown sample to all blue data samples
    - [2.5 marks] Find min_2 (minimum distance of the above distances to blue samples)
    - [2.5 marks] Compare min_1 and min_2 and assign class label to the unknown sample
- [2.5 marks] Output all unknown samples and their class label to screen
- [2.5 marks] Output all unknown samples and their class label to file
- [2.5 marks] Data sample is tuple, red, blue and unknown data samples are stored in 3 lists
- [2.5 marks] All functions are in a module file, no function is in main program
- [2.5 marks] Exception handling
- [2.5 marks] Overall
- [– 10 marks] The program cannot work with any number of dimensions
- [– 10 marks] External packages imported (except tkinter)
- [– 10 marks] Algorithm is quite different from the given algorithm
- [– 10 marks] There are no comments that explain your code

Question 2 [60 marks]

- [5.25 marks] Read data file, get number of dimensions D and number of data samples N
- [5.25 marks] Input number of clusters K, create K clusters same dimension D at random, and set threshold to a small value
- Repeat the following:
    - [5.25 marks] For each data sample, find its nearest cluster centre
    - [5.25 marks] Group data samples having the same nearest centre to a cluster
    - [5.25 marks] For each cluster, calculate new cluster centre (average of all samples)
    - [5.25 marks] Calculate sum of distances between old and new cluster centres
    - [5.25 marks] If the sum is less than the threshold: display K cluster centres and data samples on canvas then break, else: set cluster centres to new cluster centres
- [5.25 marks] Data sample is tuple, all data samples are stored in a list
- [5.25 marks] All functions are in a module file, no function is in main program
- [5.25 marks] Exception handling
- [7.5 marks] Overall (Output on canvas and Python code writing)
- [– 20 marks] The program cannot work with any number of dimensions
- [– 20 marks] External packages imported (except tkinter)
- [– 20 marks] Algorithm is quite different from the given algorithm
- [– 20 marks] There are no comments that explain your code