Dalhousie University Faculty of Computer Science
CSCI 2202 Programming For Science Fall 2022
Project 1 Due: 23 Nov 2022 11:30pm

The project problems should be solved on your own. Please refrain from using code obtained from the internet. Use a *Jupyter sheet* to do the problem Please include **text** with each problem, summarizing the results/answering questions asked. A nice way is to . The submitted code should be **modular** and **well commented**. A portion of the marks are reserved for code clarity and modularity.

(1) (a) The file `data4.txt` contains 4 data sets, with well separated means, that are mixed (randomly) into a single file. Your task is to separate the 4 data sets and report their means and std.deviations. For the separation, use the procedure below. This is an example of Machine Learning. The machine "learns the characteristics of the data sets as below:

(b) • Pick **any four** data points to use as initial seeds for the four groups.

(c) • The first four points can serve as initial centres for the $k = 4$ groups.
(You are free to select **any** *4 data points as the initial points.)*

(d) • For each data point, compute the distance to each of the four centres.

(e) • Assign the data point to the group with the *closest* centre.
(*Hint: For efficiency, use the* **squares of the distances**. *This will also help with part (f) below*).

(f) • Compute the new centroids (centres) of the four new groups generated.

(g) • Repeat steps *(a)*, *(b)*, until the centroids do not change between iterations. Use a tolerance of $1.e-4$ for this data set.

(h) • Plot the data points with **different colours** for each group at the **start** and then **after each regrouping**.

(i) • After each re-computation of the centres, mark the centre of each group on the plot of the data points.

(j) • Animate the drawing, plotting the new centroid and the new groupings as they are calculated. (the animation can be as simple like `plotLine.py` as shown in the *Root-Finding Lab* (the basis for marking) or it could be a full-fledged animation using *Matplotlib* (for bonus points).

(k) • Compute the "goodness" of the grouping by computing the sum of the squared distances of each data point in a group from its (group) centroid (RSS_k). Report $RSS = \sum_k RSS_k$. Plot RSS_k as a function of the iteration number.

(l) • Now that you have each group finalized, compute the mean and standard deviation of the corresponding x and y data for *each group*, and *plot histograms* (separately for the x and the y data for each group) of the final groupings.

(2) Download the file `iris.data` at:
`http://archive.ics.uci.edu/ml/machine-learning-databases/iris`
The dataset contains a set of 150 records of **three species of irises**, under five attributes: *petal length, petal width, sepal length, sepal width and species* (for further detail, see: `iris.name`).

Treat this as a 4-dimensional data set, using the *species* column as *labels*. In this data set, the data is labeled. The procedure remains the same:
Repeat the analyses carried out above on the **iris** data set. Identify the number of misclassified data points after applying your analysis. Once you have the groups finalized, compute the mean and std. deviation **for each of the four variables** of only one of your groups (specify the group in your solution).

Finally, identify the mis-classified points (i.e. use the labels to see if the groups are homogeneous)