Lab Test

(AI CSC3206 Semester March 2020)

Data source: https://archive.ics.uci.edu/ml/machine-learning-databases/glass/

Background: A glass can be classified into different types based on their refractive index and their composition.

The file `glass.csv` is provided. You will need to split the data into training and testing data, and evaluate the performance of a knn algorithm.

The columns in the `csv` file are:

Id	RI	Na	Mg	Al	Si	K	Ca	Ba	Fe	Glass type

- Id is imported as the row index.
- RI is the refractive index of the glass.
- Na, Mg, Al, Si, K, Ca, Ba, and Fe are the compositions of the respective chemical in the glass.
- Glass type is the type of the glass:
  - 1: building_windows_float_processed
  - 2: building_windows_non_float_processed
  - 3: vehicle_windows_float_processed
  - 4: vehicle_windows_non_float_processed (none in this database)
  - 5: containers
  - 6: tableware
  - 7: headlamps

Instructions:

1. Download the file `glass.csv`.
2. Download the file `glassKNN.py`.
3. Place the two files in the same directory.
4. The code to import `glass.csv` has been provided. You DO NOT need to write the code for that. The data is imported as a `pandas DataFrame`, saved as the variable `data`.
5. You have to achieve the following tasks:

   (a) split the `data` into 70% training and 30% testing data.

   - use Na, Mg, Al, Si, K, Ca, Ba, and Fe (*i.e.* all columns except Glass type) as the input features.
   - use Glass type as the target attribute.

   (b) plot the accuracy of knn classifiers for all odd value of `k` between `3` to `100`, i.e. `k` = `3`, `5`, `7`, . . ., `100`. This is achieved by fulfilling the following tasks:

   i. create a loop to

   A. fit the training data into knn classifiers with respective `k`.

   B. calculate the accuracy of applying the knn classifier on the testing data.

   C. print out the accuracy for each `k`.

   ii. plot a line graph with the y-axis being the accuracy for the respective `k` and x-axis being the value of `k`. You DO NOT need to save the graph.

6. Submit your code through MS Teams submission link. DO NOT upload the `csv` file and the image file of your graph. (**Note: comment out `plt.show()` if you used it in your code.** Most likely you didn't, if you are using Spyder. Don't worry if you didn't.)

Outcome:

When running your script, the script should

1. print out the accuracy for each `k` (there is no fixed format, just remember to include the values of `k` and their respective accuracy), and
2. plot the graph of accuracy against `k`.