BBM469 - Data Intensive Applications Laboratory

Assignment 2 : Clustering and Classification with Python
Date Issued : 14/03/2022
Date Due : 04/04/2022

Aim of the Assignment

The aim of this assignment is to make you familiar with clustering and classification methods using Python libraries. You will also deal with data manipulation, data normalization, and data sampling.

Background Information

We provide you with some basic tutorials for clustering and classification using Python.

Min-max normalization (from wikipedia):

Also known as min-max scaling or min-max normalization, is the simplest method and consists in rescaling the range of features to scale the range in [0, 1] or [−1, 1]. Selecting the target range depends on the nature of the data. The general formula is given as:

$$x^{'} = \frac{x - min(x)}{(x) - min(x)}$$

where x is an original value, x' is the normalized value. For example, suppose that we have students' weight data, and they vary in [160 pounds, 200 pounds]. To rescale this data, we first subtract 160 from each student's weight and divide the result by 40 (the difference between the maximum and minimum weights).

Some Useful Links

Clustering:

- https://scikit-learn.org/stable/modules/clustering.html

Classification:

- https://scikit-learn.org/stable/supervised_learning.html

Notebooks for ML

- https://github.com/krasserm/machine-learning-notebooks
- https://www.kdnuggets.com/2016/04/top-10-ipython-nb-tutorials.html

Your Assignment

1. Download the dataset. The dataset will be shared through Piazza. Choose a clustering method (Kmeans, Kmedoids, etc), and a classification method (naïve Bayes, SVM, Random Forest, etc)
2. Import and organize the dataset for clustering/classification methods (OD).
3. Normalize the dataset using min-max standardization, and create the normalized dataset (ND). Don't change the original dataset.
4. Cluster the OD dataset according to the class size of the original dataset from step 2 (set k to class size) and create the COD dataset.
5. Cluster the ND dataset according to the class size of the original dataset from step 2 (set k to class size) and create the CND dataset.
6. Present the clustering results
7. Split the datasets into training and test sets (make these steps parametric). Split the ND, OD, COD, and CND datasets with the same proportion and samples.
8. Classify the test dataset with a model trained with a training dataset.
9. Use scatter plots to show the relation between features and clusters/classes (You may use two features on two axes, and values for clusters/classes).
10. Present the classification results for each dataset (classification accuracy and confusion matrix). You will discuss the results for each task in your report with graphs and comments.

Grading

You will present your projects online sessions during laboratory hours.

- Import dataset, split training, and test dataset (%10)
- Clustering (%20)
- Visualization of clustering (%10)
- Classification (%20)
- Normalization (%10)
- Report: You will submit your report and source files before the presentation. Report details will be provided through the Piazza page (%30)

REMARKS:

- Submission format:
  - Assignment2_groupnumber <folder>
    - src.zip
    - report.pdf

- Your submission should match with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use online submission system (https://submit.cs.hacettepe.edu.tr/) to submit your experiments. Deadline is 23:59. No other submission method (such as CD or email) will be accepted.
- Do not submit any file via e-mail related with this assignment.
- The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms or source codes.
- You can ask your questions through course's Piazza group and you are supposed to be aware of everything discussed in the group piazza.com/hacettepe.edu.tr/spring2022/bbm469/home