BBM469 - Data Intensive Applications Laboratory

Assignment 2	**:** Clustering and Classification with Python
Date Issued	**:** 16.04.2020
Date Due	**:** 01.05.2020

Aim of the Experiment

This assignment aims to try to find out how we can diagnose breast cancer, using the machine learning methods from the features created by digitizing the images of breast cancer. These features define the properties of the cell nuclei in the image. Your primary purpose here is to cluster and classify the data according to the diagnosis (M = malignant, B = benign). You should be able to predict the disease most accurately by using clustering and classification methods. If there are missing values in the data set, you should explain how to deal with them. Also, you don't have to use all the features in the data set. After analyzing the dataset, you can choose which features you will use for clustering and classification.

In the first part of the study, you are expected to show the data in two separate clusters (malignant and benign) using clustering methods (Kmeans, Kmedoids, etc.). In the second part, after separating the dataset into training and test sets, you are expected to classify the test dataset with a model trained with the training dataset. As a result of the model you developed, you should be able to accurately predict whether there is breast cancer from the data obtained from a new image of the breast mass.

At the end of this exercise, you will be familiar with clustering and classification methods using Python libraries. You will also examine the data manipulation, data normalization, and data sampling approaches.

Experiment

1. Download the dataset. The dataset will be shared on the Piazza group (also, you can find explanations about the columns in the dataset next to the description file.).
2. Choose a clustering method (Kmeans, Kmedoids, etc.), and a classification method (naïve Bayes, SVM, Random Forest, etc.).
3. Import and organize the original dataset (OD) for clustering/classification methods.
4. Normalize the dataset using min-max standardization and create the normalized dataset (ND). Don't change the original dataset.
5. Cluster the OD dataset according to the class size of the original dataset from step 2 (set k to class size).

6. Cluster the ND dataset according to the class size of the original dataset from step 2 (set k to class size).
7. Present the clustering results.
8. Split the datasets into training and test sets. Split the OD, ND datasets with the same proportion and samples.
9. Classify the test dataset with a model trained with the training dataset.
10. Use scatter plots to show the relation between features and clusters/classes.
11. Present the classification results for each dataset (classification accuracy, and confusion matrix). You will discuss the results for each sub-experiment in your experiment report with graphs and comments.
12. You should submit your codes as a Jupyter notebook.
13. While grading your assignments, we will evaluate your codes with you.

Background information

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

Clustering:

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

Classification:

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

Notebooks for ML

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

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:

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

where x is an original value, x' is the normalized value. For example, suppose that we have the students' weight data, and the students' weights span [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).

Grading

You will present your projects during the laboratory hours.

- Import dataset, split the data as training and test sets (%10)
- Clustering (%20)
- Visualization of clustering (%20)
- Classification (%20)
- Normalization (%10)
- Report (%20): You will submit your report and colab code before the presentation.

REMARKS:

- Submission format:
  - studentID_name_surname_hw2<folder>
    - studentID_hw2.ipynb
    - studentID_report.pdf

- Your submission should be matched with the format above**. 10 point** penalty will be applied on mismatched submissions.
- You will use online submission system to submit your experiments.
- https://submit.cs.hacettepe.edu.tr/ 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.