Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2023

Teaching Assistants

Mohammad Ali Rezaee

Sobhan kiani

Romina Zakerian

Faezeh Hematzadeh

Assignment (5)

Outlines. In this assignment, our focus will be on unsupervised data manipulation, and we will explore some fundamental concepts in reinforcement learning.

Deadline. Please submit your answers before the end of May 30th in courses.aut.ac.ir. Other methods like sending via email or in social networks are not accepted and will not be considered.

Assignment Manual

Delay policy. During the semester, you have extra 10 days for submitting your answers with delay. Mentioned time is for all assignments. After that, for each day of delay you loss 20% points of that assignment. After 4 days you miss all points and any submit doesn`t acceptable. Remember that saving this time doesn`t have any extra point.

Sharing is not caring. Students are free to discuss and share their ideas about problems with others. But sharing source codes, solutions, answers and other results is not allowed and based on university`s rule, both sides will be graded zero.

Problems are waiting you. Some problems are required to be implemented within a programming language and obtain some charts, images, results, etc; then discuss about it. These types of questions are tagged by #Implementation. Some other problems are required to be solved or computed by hand or research about them. These types of questions are tagged by #Theorical. You are not allowed to use programming language or other technical tools to answer theorical problems.

Report is the key. All students' explanations, solutions, results, discuss and answers must be compacted into a single pdf report. A clean and explicit report is expected and may followed by extra pts; so, you may need to write any related detail or experience during the solving problems. Report file should started within a cover page that it includes course and assignment information as well as identical details like name, student number and email address. Second page should be table of contents that indicates student`s answer to each question. Please repeat your name and student number in left side of footer in other pages. Also, you are free to write in Persian or English. If typing is bothering you, so write in a paper and put its picture with acceptable readability quality in report file.

Organize the upload items. Students should upload their implementation source codes as well as results and report. You should upload a single .zip file with the following structure:

ML_05_[std-number].zip
- Report
  - ML_5_[std-number].pdf
  - [other material and results]
- Source codes
  - P[problem-number]_[a-z].py
  - P[problem-number]_[a-z].ipynb
  - …

Python is the power. Students are free to use any programming language like python, matlab, C++ , etc. However it is recommended strongly to use python in jupyter notebook environment; so, you may need to upload your .py or .ipynb sources.

Feel free to contact. If you have any question or suggestion, need guide or any comment be comfortable to ask via email as well as Telegram group. Good luck with your learning journey!

Problem 1: **(24 pts)**

A group called HELP, which is not affiliated with any government, plans to direct its aid to the countries in greatest need. To do this, they've shared a dataset with information on various social, economic, and health factors, to assess the development levels of different countries. The goal is to cluster the countries based on this data and identify those most in need of aid.

The data columns are as follows:

- **country:** The name of the country.
- **child_mort:** Deaths of children under 5 years of age per 1000 live births.
- **exports:** Exports of goods and services as a per capita .Given as %age of GDP per capita.
- **health:** Total health spending per capita .Given as %age of GDP per capita.
- **imports:** Imports of goods and services per capita .Given as %age of GDP per capita.
- **Income:** Net income per person.
- **Inflation:** The annual growth rate of the total GDP.
- **life_expec:** The average number of years a newborn would live if current mortality patterns remain the same.
- **total_fer:** The number of children that would be born to each woman if current age-fertility rates remain the same.
- **gdpp:** GDP per capita calculated as the total GDP divided by the total population.

1. Examine the correlation matrix to assess the correlation between features. Determine if some features can be removed and explain the rationale behind it.

2. Evaluate whether the data needs normalization. Justify your response and perform the necessary operations on the data.

3. Clustering:
   - Use the Elbow method to determine the optimal number of clusters in the k-means algorithm and present the result.
   - Research and briefly explain the silhouette score, a metric for evaluating the quality of clustering.
   - Determine the optimal number of clusters using the silhouette score and compare it with the Elbow method results. (You can use the `silhouette_score` function from the sklearn library for implementation).

4. Execute the k-means algorithm with the optimal number of clusters and provide the cluster number for each data point.

5. Select three features of your choice and plot them in scatter diagrams, showing each cluster in different colors. Interpret the results.

In this section, the goal is to reduce the dimensionality using the PCA algorithm and identify the principal components.

1. Execute the PCA algorithm on the normalized data (You can use the PCA function from the sklearn library for implementation).

2. Determine how many principal components can adequately explain the data distribution. Use a Percentage of Variance Explained by the Components chart to present your findings and discuss the decision-making criteria in your report.

3. Reduce the dataset dimensions based on the results (retain the essential components and discard the rest).

4. Perform clustering on the reduced data using the k-means method as described in steps 2 to 4 above. Compare and interpret the results with the previous clustering outcomes.

Problem 2: **(16 pts)**

In this exercise, we intend to familiarize ourselves with the performance of perceptron and Adaline neural networks in data classification. To perform this exercise, first, the data generation process is explained. Then, answer the following questions.

1. Generate a two-dimensional linearly separable dataset in two classes. To generate this dataset, use a Gaussian distribution with different means and identical covariance matrices. **Explain** how you would set the values of the means and the covariance matrix of this distribution to ensure that the generated dataset is linearly separable. Using the described process, generate a two-class dataset with 10,000 data points (5,000 from each class) and plot the distribution of the data.

2. Design a perceptron neural unit and an Adaline neural unit for classifying the generated dataset in question 1. **Explain** the structure of the input, initial values of weights, and the complete activity of these neurons. What is the difference between these neurons?

3. Train each of the above neurons for the classification task on the dataset. During the training, plot the training and validation error graphs, **report** the convergence speed, and the final accuracy on the test set. Compare the performance of these neurons in classifying the dataset and state which of these neurons you think is more suitable for this classification task.

Problem 3: **(16 pts)**

Please review the Jupyter notebook titled "RL_HW5" and fill in the missing sections of the code. Additionally, prepare your report.

Problem 4: **(16 pts)**

In this exercise, we have images of handwritten digits from 0 to 9. We aim to use a clustering algorithm to put similar digits in the same cluster.

a) Load the dataset using load_digits method of sklearn library. Explore the structure of the dataset. Then for each digit plot at least one image.

b) If needed, perform necessary pre-processings on the dataset.

c) First Use KMeans algorithm to create clusters from the digits.

d) Second, Use a Hierarchical clustering algorithm to detect the clusters.

e) Using proper metric(s) evaluate the quality of the models trained in parts b and c.

f) We want to assign a label to each cluster of digits. What is your approach to do that? Write the code for it.

Problem 5: **(12 pts)**

In this exercise we aim to detect clusters in some datasets using KMeans and DBSCAN algorithms.

a) First, load the datasets "Dataset1.txt" and "Dataset2.txt". Plot each of the datasets and get familiar with them. What is your suggested number of clusters for each of the datasets?

b) Train KMeans clustering on the datasets.

c) With proper metric(s), evaluate the quality of clusters. Then use elbow method to find the optimal number of clusters. Report and analyze your results for each dataset.

d) Train DBSCAN clustering on the datasets. Find the best parameters to get the best possible result for the datasets. Then briefly explain the chosen parameters.

e) With proper metric(s), evaluate the quality of DBSCAN clusters. Compare your results with the results in part c. Can you draw any conclusions from these results?

f) How can we detect outliers using KMeans algorithm? Load the "Dataset3.txt". First, plot the dataset to get familiar with data. Then, use KMeans clustering algorithm to detect outliers. Plot the results by showing the outliers with a different color.

Problem 6: **(16 pts)**

Quantization refers to a method in which we express a range of values with a single value. For images this means that we can compress the entire color spectrum into a specific color. This method is lossy, i.e. we intentionally lose information in favor of less memory consumption. In this exercise, implement the quantization method using the K-means clustering. In better words, Convert the image “stones.jpg” to step mode with values of 1, 2, 4, 8, 16, and 32 using k-means clustering and show the results.