Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2023

Teaching Assistants

Mohammad Ali Rezaee

Amirhossein Kashani

Romina Zakerian

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 January 2th 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: Practicing the Fundamentals of Clustering (12 pts)

#Theorical

answer the below question.

a) Select True of false then describe K-means always converge to the local optimum (true or false) Any initialization of the centroid in k-means is just as good as just any others (true or false) DBSCAN and k-means both can detect curved clusters (true or false)

b) Explain the difference between k-means and k-means++. does initializing centroid using k-means++ guarantee convergence to global optimum?

c) Suppose we initialize k-means with the following structure (figure 1). Draw the next position of each centroid in the next steps until stabilizing.

d) Consider the 2D dataset below and apply the hierarchical-based clustering with a single link approach; indicate all calculation steps and plot the dendrogram.

	#1	#2	#3	#4	#5	#6
X	34	34	25	14	13	18
Y	14	6	9	6	8	6

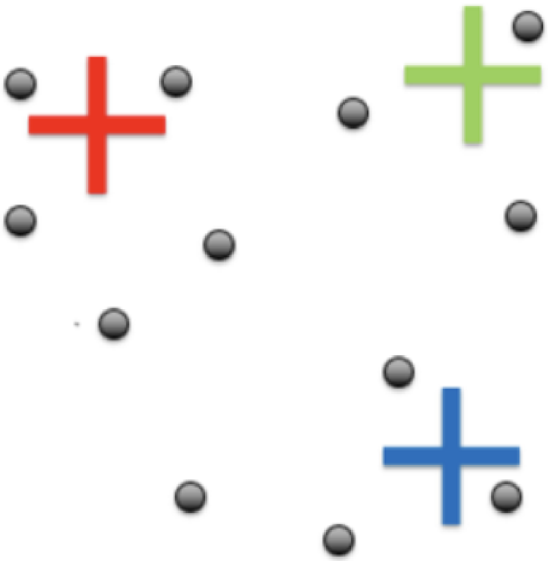


Figure 1

Problem 2: Understanding the Bellman Equation in Reinforcement Learning (8 pts)

#Theorical

The Bellman equation, named after Richard Bellman, helps us solve the Markov decision process (MDP). When we say solve the MDP, we mean finding the optimal policy. the Bellman equation is ubiquitous in reinforcement learning and is widely used for finding the optimal value and Q functions recursively. Computing the optimal value and Q functions is very important because once we have the optimal value or optimal Q function, then we can use them to derive the optimal policy.

a) Your task is to elucidate the Bellman equation for both the value function and the Q function in both stochastic and deterministic environments. Please present the equations and furnish a comprehensive explanation. Additionally, if the action space is stochastic, please expound on how the equations are modified to accommodate this stochasticity.

b) How do we derive the value function from the Q function?

c) How do we derive the Q function from the value function?

Problem 3: Clustering Using Representatives (20 pts)

# Implementation

CURE is an efficient data clustering algorithm for large databases. Compared with K-means clustering it is more robust to outliers and able to identify clusters having non-spherical shapes and size variances. In This question, we want to implement the CURE algorithm and cluster two moons data sets.

a) First load the dataset with this command:
   a. from sklearn.datasets import make_moons
   b. data = make_moons(n_samples=10000, noise=0.05)

b) Pick 500 random sample data

c) By using hierarchical clustering find the two clusters of data (in this section you can use the library)

d) Plot the clusters using a scatter plot with different color markers for each cluster.

e) Now randomly select m point from each cluster we call these point representation points. Plot the points

f) Calculate the center of each cluster based on the representation points and shift each representation point 10 percent closer to the center of its corresponding cluster.

g) Now for all 10000 data points find which clusters by finding the nearest representation point.

h) Plot the final result.

i) Mention the time complexity of CURE algorithm.

Problem 4: Breast Cancer Prevention using K-Means Algorithm (20 pts)

# Implementation

Breast cancer, a prevalent form of cancer globally, accounted for 12.5% of new cases in 2020. Despite its severity, early detection through regular screenings and tests is possible. Fine Needle Aspiration (FNA) is an effective method for identifying breast cancer, involving the extraction of a small tissue sample using a syringe, followed by imaging. Clinicians then aim to isolate individual cells in each image to extract 30 characteristics, including size, shape, and texture. Our objective is to utilize K-Means clustering to diagnose breast cancer based on these features extracted from the Wisconsin Diagnostic Breast Cancer dataset.

a. Define a Python class, KMeansCluster, to implement the K-Means clustering algorithm. Include an initialization method that accepts parameters for the number of clusters (k), convergence tolerance (tol), and maximum iterations (max_iter).

   def __init__(self, k, tol, max_iter)

   Implement functions within the class for fitting the model to data (**fit**), calculating clustering accuracy (**accuracy**), predicting cluster labels (**predict**), and computing the sum of squared errors (**sse**). The fit method should use the input data matrix (X) and matrix of initial centres (mu) to perform K-Means clustering iteratively. The clustering result is also stored in matrix C. Use your class to cluster the data, and report the accuracy of the clustering.

b. Run your code 5 times using different starting points, and calculate the accuracy of each case. What were your observations? Explain.

c. Run your code using the provided initial centres (init_mu (in dataset folder)), in which each column represents one of the initial centres, and report the accuracy of the clustering.

d. What happens if you initialize with the true centres, obtained after the true clustering?

e. Can you achieve better accuracy using another unsupervised learning method? What about a supervised one? Explain and implement.

Problem 5: Image Compression (25 pts)

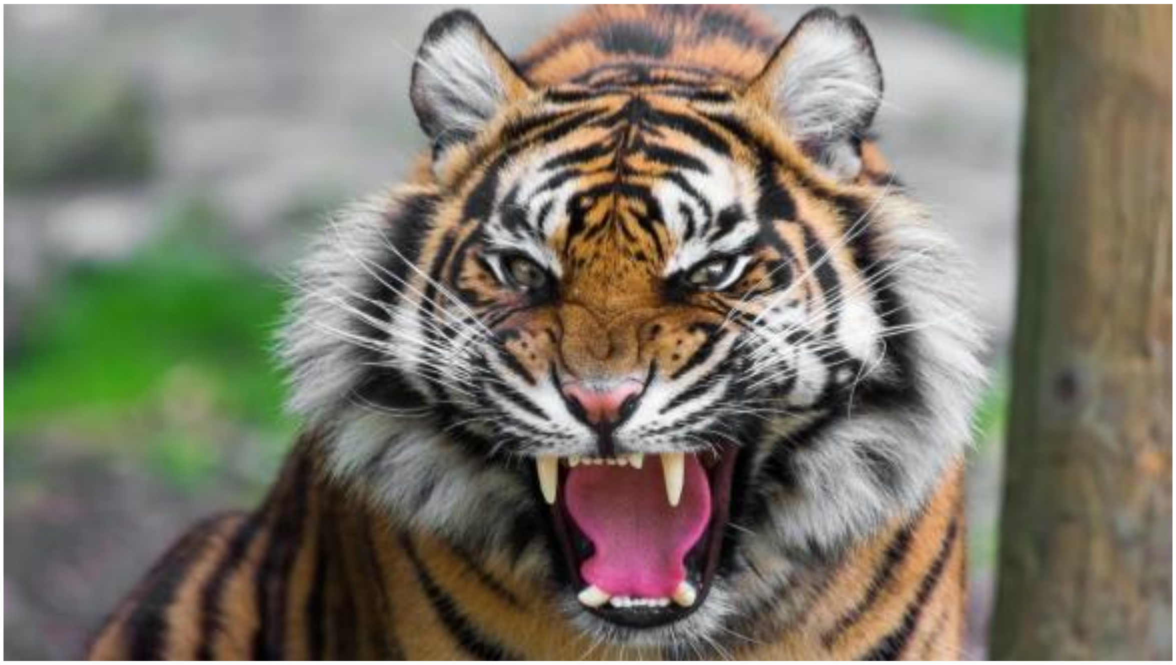
# Implementation

The internet is filled with huge amounts of data in the form of images. People upload millions of pictures every day on social media sites such as Instagram, and Facebook and cloud storage platforms such as google drive, etc. With such large amounts of data, image compression techniques become important to compress the images and reduce storage space. In this article, we will look at image compression using the K-means clustering algorithm which is an unsupervised learning algorithm. An image is made up of several intensity values known as Pixels. In a colored image, each pixel is of 3 bytes containing RGB (Red-Blue-Green) values having red intensity value, then Blue and then green intensity value for each pixel.

Approach: K-means clustering will group similar colors together into 'k' clusters of different colors (RGB values). Therefore, each cluster centroid is representative of the color vector in the RGB color space of its respective cluster.

Your task is to compress the **tiger** image with k-means clustering and then show the output image in the doc.

In the doc, explain the method and steps took for this task.



Problem 6: More Into Clustering (15 pts)

# Implementation

In this section, our objective is to implement the DBSCAN class for clustering tasks and apply this algorithm to the datasets "pathbased-D31-spiral-Compound." The requirement is to refrain from using any library that implements DBSCAN. After completing the implementation, we will address the following questions:

a) Calculate the purity criteria and determine the number of clusters achieved using the implemented DBSCAN class.

b) Visualize and show the data related to each cluster with a different color. Note that some data may not belong to any cluster and may be considered noise. Show noisy data with different color.

c) What effect does the types of datasets have on the performance of the DBSCAN algorithm?