Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2022

Teaching Assistants

Mohsen Ebadpour(head) (M.Ebadpour@aut.ac.ir)

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Mohamadreza Jafaei (Mr.Jafaei@aut.ac.ir)

Assignment (5)

Outlines. In this assignment, Unsupervised and Reinforcement learning are noticed.

Deadline. Please submit your answers before the end of January 23rd in courses.aut.ac.ir. Other methods like sending via email or in social networks are not accepted and will not be considered.

NOTE: Latest time that you can upload your answer with delay is the end of January 27th.

Assignment Manual

Delay policy. During the semester, you have extra 7 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_05_[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.

Problem 1: why and how (35 pts)

#Theorical

a) Any clustering method or algorithm can be used for outlier detection. With supporting examples, Describe the strategies which are behind it.

b) Explain the Expectation-Maximization(EM) algorithm and indicate that K-Means is the specific form of EM, which consist of K Gaussian functions. (Discuss their covariance)

c) Consider the figure below; we repeatedly applied the EM and K-Means clustering algorithm with random initialized starting points, in which the cluster count is two, and picked the best results. With suitable supporting statements, show the centers of two clusters in the figure.

d) We want to cluster the customers of a shopping center for advertising goals. Which kind of approach is the better choice? Explain your reasons. (partitioning-based, density-based, hierarchical-based, and model-based)

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

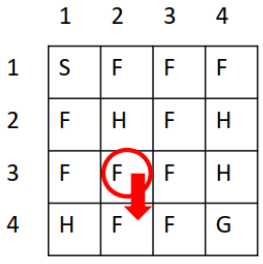
ID	*{1}*	*{2}*	*{3}*	*{4}*	*{5}*
X	14	18	25	34	34
Y	6	6	10	6	14

f) Consider the 1D dataset below, apply the K-Means algorithm with two partitions, and update the centers until convergence. (init the centers randomly with the calculator in range -20 and +20) What will happen if two init centers happen in the range of 15-20? Discuss this situation and talk about its solutions.

ID	*{1}*	*{2}*	*{3}*	*{4}*	*{5}*	*{6}*	*{7}*	*{8}*
X	-7	+10	+7.5	-6	-12	-9	-11	+4.5

g) Describe the Gaussian Mixture Model (GMM) and its hyperparameters. What are the advantages and disadvantages; For what kind of datasets is suggested? justify your answers.

h) [3 extra pts] Consider the simple example of a frozen lake. the agent is in state (3, 2). Moving to F state has a -0.3 reward. The learning rate is 0.1 and the discount rate is 0.9. According to the Q-table, if the agent chooses the down action, update the Q-table.



Q table

State	Action	Value
(3, 2)	Left	0.1
(3, 2)	Right	0.5
(3, 2)	Down	?
(4, 2)	Up	0.3
(4, 2)	Down	0.5
(4, 2)	Right	0.8

Problem 2: Blood Analysis | HCV (30 pts)

Implementation

In this problem, we want you to cluster the blood tests to define whether its owner can be a blood donor:

a) Load the dataset and apply the needed preprocessing steps. (Categorical to numerical and normalization). why is normalization in distance-based clustering algorithms necessary? Justify your answer with examples.

b) Implement the K-Means function with the below template:
KMeans(X_train, ClusterCount=2, iteration=30, ConergenceRatio=0.01, DistanceMetric=[Euclidean or CityBlock]) -> centers: list

c) Use the implemented function to detect outliers in the dataset and indicate what percent of the dataset is recognized as an outlier; Explain the steps you have done. (Note that the target column is "Category" and should be removed in the feature vector)

d) Split it to train and test sets with the ratio of 8:2.

e) Train the K-Means clustering algorithm for two centers (target to evaluate [0-0s] or [1-3]) and test it with obtained centers and test sets. Report the results for Purity, Entropy, and accuracy metrics as well as obtained centers. (Note that labels of obtained and original may be different only numerically; for example, the original label of 0,1 is assigned to 1,0 respectively, and there may be no difference in members) You are not allowed to use functions or libraries to calculate metrics.

f) Repeat the previous part for four clusters (target to evaluate: [0-0s], [1], [2], or [3]).

Problem 3: DBScan is talking(35 pts)

Implementation

In this problem, you are asked to play with some datasets and get more familiar with DBScan clustering algorithm hyperparameters:

a) Load each of the given datasets and plot them. Based on your observation, discuss the count of clusters in each dataset and outliers.

b) Implement the DBScan function with the below template:
DBScan(X_train, Epsilon, MinPoints) -> labels : list(its length is equal with X_train)

c) With sufficient tries and using the above function, cluster each dataset and plot the clusters (each with unique color).

d) Implement the function below, which makes it possible to predict new samples based on obtained clusters:
DBScanPredict(x_train, dbscan_cluster_labels, x_sample) -> predicted label for x_sample
Describe your strategy for assigning cluster labels based on DBScan.

e) For each dataset, split it into 85:15 train and test sets. Predict the test samples using DBScanPredict based on obtained clusters via the DBScan. Report the accuracy, discuss results, and plot test samples to indicate that assigned to which cluster.(with colors)

f) Now, train the 3-NN and 5-NN models for each data set with the above situation, calculate accuracy, and compare results with the previous part. Which one is better? Why? Justify your answers.

Problem 4: Frozen Lake (Extra 25 pts)

Implementation

Gym[1] is an open-source Python library for developing and comparing reinforcement learning algorithms by providing a standard API to communicate between learning algorithms and environments, as well as a standard set of environments compliant with that API. Consider a frozen lake environment. Frozen lake involves crossing a frozen lake from Start(S) to Goal(G) without falling into any Holes(H) by walking over the Frozen(F) lake. The agent may not always move in the intended direction due to the slippery nature of the frozen lake.

a) Describe this environment. (Action Space, Observation Space, Rewards end etc.).
b) Run an agent using a random action for 1000 episodes (any episode length). How many times did the agent win?
c) Train an agent for this environment (4*4) using Q-learning for 100,000 episodes.
d) Report the final Q-table and final policy.
e) After extracting the final policy, test it for 1000 episodes and compare it with part b of the question where your policy was random.
f) Render a video gif for one episode that agent wins.

[1] **https://www.gymlibrary.dev/**