Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Spring 2024

Teaching Assistants

Sobhan Kiani (sobhankiani@aut.ac.ir)

Mehdi Hosseini (mehdi.hsn@aut.ac.ir)

Donya Haddad (donya.haddad@aut.ac.ir)

Amir Masoud Sepehrian

Assignment (1)

Outlines. In this assignment, some practical implementation skills which needed in this and other courses of this degree are noticed as well as regression topics. Remember that you may need to re-use your implementations of this assignment; so, it is suggested to code in functional.

Deadline. Please submit your answers before the end of date 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 5 days for submitting your answers with delay. Mentioned time is for all assignments. After that, for each day of delay you lose 20% of the points of that assignment. After 4 days you miss all points and any submission will not be 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 for you. Some problems are required to be implemented within a programming language and obtain some charts, images, results, etc; then reasoned about. These types of questions are tagged by #Implementation. Some other problems are required to be solved or computed by hand or researched about. These types of questions are tagged by #Theoretical. You are not allowed to use programming language or other technical tools to answer theoretical problems.

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

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

```
ML_01_[std-number].zip
        Report
                ML_01_[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 strongly recommended to use python in the jupyter notebook environment; so, you may need to upload your .py or .ipynb sources.

Feel free to contact us. If you have any question or suggestion, need guidance or any comment be comfortable to ask via email as well as Telegram group.

Problem 1: Table Preprocessing (40 points)

In machine learning, preparing data for model training typically involves several preprocessing stages. These steps, collectively referred to as preprocessing, include tasks such as data cleaning and feature engineering. While some techniques are commonly applied across various datasets, others are employed for specific cases. This exercise focuses on the essential preprocessing steps commonly undertaken before model training.

a. Load CSV file of the dataset “Dataset_01.csv” and show 10 random items.

b. Visualize the data to find more about the characteristics of each feature.

c. Visualize the data to find the relationship between different features specially features with the target value.

d. Perform necessary data cleanings (which includes handling missing values, duplicates, outliers, inconsistencies, incorrect values, etc.).

e. Perform the necessary feature engineering and normalize the data.

f. For each step, describe your approach and the results.

Problem 2: Image Preprocessing (20 points)

In this section, we want to focus on some of the fundamentals of image preprocessing. Image processing is not as easy as it may seem, and numerous techniques may be performed on an image before being used as data in a machine learning model. Follow the steps below to perform some basic processing on the photos.

a. Load the images int the 'Dataset_02' directory using the 'opencv' library.
b. Select 3 random images and display them on the screen. Explain the three dimensions in the shape of the images.
c. What are the pros and cons of using color images compared to grayscale images? Apply grayscale conversion to the images if needed.
d. What are the effects of image brightness and contrast on the visual quality and image interpretation? Why would you consider adjusting these parameters as a preprocessing step? Adjust the brightness and contrast of the images using proper approach(es).
e. What is image normalization? What challenges may arise if images are not normalized? properly normalize the loaded images.

Problem 3: Text Analysis (40 points)

In this section of the assignemnt, we are going to see the preprocessing steps taken to process text. Hamshahri newspaper is a great dataset to work on.

a. Explore and experiment with the dataset. Get the dataset using the following link: https://drive.google.com/file/d/1D3yt99D0GcCRCbdKbUQGxbqjkeh91hTg/view
b. Read the data from the "HamshahriCorpus.txt"file and convert the data into a tabular format.
c. Visualize the data using different plots. What can be understood from each plot?
d. What preprocessing steps would you take to clean this dataset? Apply necessary preprocessing steps.
e. Extract useful information from the text using methods such as Term Frequency-Inverse Document Frequency (TF-IDF).
f. Find five of the most important and frequently used words in Hamshahri Corpus.