Amirkabir University of Technology
(Tehran Polytechnic)

Applied Machine Learning Course By

Dr. Nazerfard

CE5501 | Spring 2023

Teaching Assistants

Mohamadreza Jafaei (HEAD) (Mr.Jafaei@aut.ac.ir)
Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)
Ali Amirian (ali.amiryan@aut.ac.ir)
Ghazaleh Gholinejad (Ghazaleh.gholinejad@aut.ac.ir)

Project

Outlines. In this project, All Topics are noticed (3 Questions).

Deadline. Please submit your answers before the end of July 12nd 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:

```
AML_01_[std-number].zip
        Report
                AML_Project_[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: **Price Predictor:** Web Scraping and Machine Learning for Peugeot 206 Type 2 **(35+5 pts)**

#Implementation

In this question, we ask you to train a model to predict the price of Peugeot 206 type 2 using machine learning. But where can we get the price data of Peugeot 206 type 2? In this question you will learn how to do web scraping. For this purpose, follow the steps below.

1- Divar, a popular website, publishes numerous car sales ads daily. We can extract valuable information from these ads to create the dataset we need. Selenium library can be used for web scraping.

2- By analyzing the car advertisements, we can identify important factors that can affect the car's price, such as the year, mileage, color, etc. It would be beneficial to extract and save these important features. Additionally, descriptions and photos of the cars can also provide important information (bonus points).

3- Next, perform the required pre-processing and train the model.

4- Evaluate the model and report.

5- Finally, we need to save the trained model and develop a script that takes car information as input and predicts whether the price suggested by the seller is reasonable or not.

Problem 2: Hamshahri Newspaper (30+10 pts)

# Implementation

In this project, we intend to work on the Hamshahri newspaper dataset, step by step, to fully implement the project and explain all the steps in our report.

Step 1 - Introduction to the Dataset

a) Please visit the following link and familiarize yourself with the dataset. In a paragraph, provide a general description of the dataset. https://dbrg.ut.ac.ir/hamshahri

b) Download the dataset using the following links:

- Direct Link: http://dbrg.ut.ac.ir/wp-content/uploads/datasets/hamshahri.rar

- Google Drive Link: https://drive.google.com/file/d/1D3yt99D0GcCRCbdKbUQGxbqjkeh91hTg/view

Step 2 - Data Loading

a) Read the data using your preferred method. In this section, we will work with the file "Hamshahri-Corpus.txt," so you need to have access to that file. The image below shows the beginning of the file.

b) Convert the text file into a table format, as shown in the example image.

Step 3 - Data Visualization

a) Visualize the data using three different types of plots.

b) Describe each of the generated plots.

Step 4 – Preprocessing

a) Perform the necessary preprocessing steps on the text.

b) Explain the purpose of each preprocessing step.

Step 5 - Feature Engineering

a) Extract useful information from the texts using methods such as Term Frequency-Inverse Document Frequency (TF-IDF).

b) Calculate the frequency of the word "ناتو" (NATO) at different time intervals from the beginning until now and plot the results.

c) Analyze the plotted graph and explain the reason for the maximum value by conducting a web search.

Step 6 - Dimensionality Reduction

a) Reduce the dimensionality of the data to 2 dimensions using dimensionality reduction tools.

b) Plot the data based on the two obtained dimensions.

Step 7 - Clustering In this section, consider that the category should not be treated as data or information (unsupervised learning).

a) Perform clustering operations based on the data, considering the number of clusters you have determined.

b) Visualize the clustering results using the given labels.

c) Compare this plot with the plot from Step 6, Part b, and provide an explanation.

Step 8 - Storage

a) Perform dimensionality reduction again, but this time change the number of dimensions to 5.

b) Apply the clustering algorithm on the obtained values and save the output labels.

c) Save the file as the final result.

Step 9 - Classification (Model Building)

a) Consider the models: KNN, Logistic Regression, Naïve Bayes, and Random Forest.

b) Create an ensemble model called "ensemble_model" that combines the results of the above five models through voting.

Step 10 - Preprocessing on Data

a) Preprocess the data.

b) Split the data into training and testing sets with a ratio of 1 to 4.

Step 11 - Model Training

a) Train the models.

Step 12 - Model Evaluation

a) Calculate the accuracy for each model separately on the training and testing data.

b) Obtain the Confusion Matrix based on the testing data.

c) Analyze the outputs.

d) Which model performed better compared to the others? What is the reason for its better performance?

e) If we used the Linear Regression model for this task, what would be the result? Explain.

Step 13 - API (Extra Credit)

a) Initially, make sure that any label assigned to you through clustering is mapped to the best original category.

b) Implement an API using your preferred library to classify a given text and return its category.

Problem 3: Face Recognition (35+5 pts)

# Implementation

In this question, we are going to use the traditional networks that we discussed in the course to recognize the face. You need to design a system that receives the incoming image from your computer's webcam and prints the message hello if your face exists in the image and prints the message “can’t login” if the face of anyone else but you exist in the image.

For this question, you need to collect a data set only from your face. Then, using the traditional networks that you learned in the course (it is not possible to use the neural network), train a model to compare the input image from the webcam with your images, and if the similarity is higher than a limit, it will give you the possibility of entering. (In this section, attractive ideas with high accuracy are given extra points.)

At the end of the program, after starting 20 consecutive forms, it should receive the webcam image and if it recognizes your face in a certain number of these images, it will allow you to enter.

You can use cv2 to receive video frames from webcam, and if your system does not have a webcam, you can create a virtual webcam on the system using **many cams** software and using your phone.

The file sent for this exercise must include the video and collected data so that it is possible to test and run the code. If you don't want your own images to be included in the delivered file, replace your data file with the images of a famous Iranian artist and display the artist's image on your phone to simulate the result and bring it close to your system's webcam. Just note that the code delivery day is evaluated based on your face.

Please note that the proposed method must be able to determine the result of entry or non-entry in an acceptable time, and the output of this code will be checked, and your code must be able to be executed on the day of submission.