Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Spring 2024

Teaching Assistants

Amirmasoud Sepehrian

(amirmasoud.sepehrian@aut.ac.ir)

Poorya Azizi (poorya.azizi@aut.ac.ir)

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

Nima Hatami

Assignment (2)

Outlines. In this assignment, some practical implementation skills which needed in this and other courses of this degree are noticed as well as linear regression and KNN 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_02_[std-number].zip
    Report
        ML_02_[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: California Housing Prices | Regression (23 points)

The purpose of this question is to train a polynomial regression to estimate housing prices. The attached data is related to the average house price in different areas of California. This data includes 8 numerical features and one categorical feature. The number of samples is 20640. And the house prices are located in the value_house_median column. For this purpose, do the following steps (all the steps below should have a report):

a. Load CSV file of California Housing Prices and show 10 random items
b. Data visualization (at least 5 plots with interpretation)
c. Perform the necessary data cleanings such as handling missing values, duplicates, outliers, incorrect values and etc
d. Perform the necessary feature engineering, encoding, scaling and train/test/validation split
e. Train the model with k-fold cross validation in order to find the best degree. Draw a plot to show the changes in MSE on the validation data with increasing regression degree
f. Train the final model with best degree and report MSE and R2 on the test data.
g. Calculate the following two features for each sample and add them to the dataset:
    - Population_per_household = $\frac{population}{households}$
    - Rooms_per_household = $\frac{total\_rooms}{households}$

   Perform the steps d to f again

h. What did you deduce from the previous step? explain your opinion

Problem 2: Predicting 10-Year CHD Risk | KNN (23 points)

The goal of this question is predicting a patient's 10-year risk of coronary heart disease (CHD). The dataset contains information from an ongoing cardiovascular study in Framingham, Massachusetts.

a. Load dataset " Framingham" and provide information about the features and the number of their missing values.
b. Apply the required pre-processing on the data. including removing duplicate rows, handling missing values, etc.
c. Using SelectKBest from the sklearn library, select 10 best features that have larger contribution towards the outcome variable, TenYearCHD. In one paragraph explain how SelectKBest works.
d. Split the dataset into training and testing (40/60) and perform data scaling.
e. Train the data using the KNN algorithm with k=3 and classify it.
f. Obtain the Confusion matrix, R2_Score, and Accuracy.
g. Select a different distance metric of your preference and repeat steps e and f. Elaborate on the impact of the chosen distance metric on the results. Is it effective in improving performance?
h. Perform additional preprocessing methods to improve results. *(extra point)*

Problem 3: Spam Detection | KNN (19 points)

This section of assignment focuses on detecting Persian spam emails using machine learning algorithms. The goal is to develop an effective spam detection system using KNN.

a. Load CSV file of "email" Dataset
b. Perform the necessary text preprocessing and provide a rationale for each step. (You can use the hazm library, which is used to process the Persian language).
   (Note that this section has the highest score).
c. Split the dataset into training and testing (70/30) and perform data scaling.
d. Train a KNN classifier on the training set using different values of K ranging from 1 to 20, and evaluate the performance of the classifier on the test set by computing the classification accuracy for each value of K. Choose the optimal value of K based on the accuracy results.

Problem 4: Image Classification | KNN (35 points)

The Cats-vs-Dogs dataset is designed for binary classification, The dataset includes both test and train data, divided into two folders. The training data includes 12,500 images of dogs and 12,500 images of cats along with their correct class in the image title (e.g. dog-8837.jpg). The test data includes 12,500 images of dogs and cats without their correct classifications (e.g. 1.jpg). You should do this classification using KNN. Do the following steps:

a. Download and Load Dataset. (Download dataset from here)
b. Pre-processing: The size of the images may be different. Resize them. Then normalize the images. Determine the label of each image. (We suggest using glob and OpenCV libraries).
c. Split data to train and test with 80/20 ratio.
d. Train a KNN classifier on the training set using K=5 and the default distance metric. Evaluate the performance of the classifier on the test set by computing the accuracy, precision, recall, and F1-score. You can do it with 'classification_report' function.
e. Use 10-fold cross-validation to evaluate the performance of the KNN classifier with K=5 and the default distance metric. Compute the average accuracy, precision, recall, and F1-score, along with their standard deviations.
f. Is KNN suitable for image classification? Why?

Note: If you are encountering constraints such as limited processing resources or time, it is advisable to utilize a subset of the dataset. (Perform dataset sampling on step 'a').