Amirkabir University of Technology
(Tehran Polytechnic)

Applied Machine Learning Course

By Dr. Nazerfard CE5501 | Spring 2024

Teaching Assistants:

Romina Zakerian

Mehdi Hosseini

Donya Haddad

Mohammad Ali Rezaee

Amir Hossein Babaeayan

Assignment (4)

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 4 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_04_[std-number].zip
- Report
  - ML_04_[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: Data Collection and Web Scraping (25+10 points) - AmirHossein

---

1. **Data Collection and Web Scraping**:

    - Data is a critical component in artificial intelligence and machine learning. Collecting data from various sources can be costly and time-consuming. One method for data collection is web scraping, which involves automatically retrieving data from websites.
    - The repository you mentioned, divar-web-crawler-BS4, demonstrates web scraping using BeautifulSoup (BS4). It collects information related to the Peugeot 206 sedan from the "Divar" website. The collected data is then used to build a linear regression model after preprocessing.
    - Keep in mind that this repository may not represent the best performance, and you can explore other libraries like Selenium for web scraping.

2. **Tasks to Complete**:

    - **a) Familiarize Yourself with the Repository**: Study the code in the repository to understand how web scraping is implemented.
    - **b) LinkedIn Profile Setup**: If you're not familiar with LinkedIn, take the time to create an account, complete your profile, and explore the platform.
    - **c) Implement a Web Scraper**:

        - Create a web scraper that collects data on active individuals in the "Software Engineering" field in Iran. These individuals should currently work at one of the top 50 domestic companies.
        - The dataset should include at least the following features for each person:

            - Full name
            - Email (if available)
            - Website (if available)
            - LinkedIn headline
            - City and country of residence
            - Current company
            - Three work experiences
            - Three educational qualifications

            - Five key skills

        - Aim for a dataset with at least 2000 records.

- **d) Prepare a Program**:
    - Develop a program that can generate a dataset with a minimum of 100 records based on the following input criteria:
        - Field of activity
        - Country
        - University
        - Current company
- **e) Data Preprocessing Suggestions** (bonus points):
    - To clean the data effectively, consider the following preprocessing steps:
        - Handle missing values (e.g., impute or remove them).
        - Standardize or normalize numerical features.
        - Encode categorical features (e.g., one-hot encoding).
        - Remove duplicates.
        - Address outliers.
        - Tokenize and preprocess text data (e.g., LinkedIn headlines).
        - Explore feature engineering opportunities (e.g., creating new features from existing ones).

1. AmirHosseinBabaeayan/divar-web-crawler-BS4: with BS4 - GitHub
2. divar-web-crawler-BS4 - GitHub
3. Releases · AmirHosseinBabaeayan/divar-web-crawler-BS4 - GitHub

Problem 2: Airplane crash (30 points) – Donya and Mehdi

There has been an airplane crash recently. A lot have lost their lives, but some have survived. While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

The dataset provided along with the assignment document contains passenger records. We need you to build a classifier that could helps us answer the question: "what sorts of people were more likely to survive?"

a) Load the dataset and prepare it for model training.

b) Split the data into 80% train and 20% test dataset.

c) Train an SVM model and report accuracy, precision and recall scores.

d) Use grid search with cross validation to extract the best hyper parameters.

e) Report your findings along with the reason behind hyper parameter value.

f) Train another SVM model with the best hyper parameters and report accuracy, precision and recall.

Problem 3: Credit Card (25 points) – Mohammad Ali

Some credit card companies have designed methods using algorithms that, based on purchase data, determine whether a purchase was made by the original cardholder or if it is fraudulent. The dataset is located at:
https://drive.google.com/file/d/1sq18MxrJKh1KbjeSo3OKzmlhD3puUD37/view?usp=sharing

This dataset contains information about purchases made by citizens over two days. If the purchase is valid, a zero is placed in the last column; otherwise, a one is placed.
In total, there were 284,807 valid purchases and 492 invalid purchases. Columns v1, v2,... represent the values of variables obtained after dimension reduction using PCA. We intend to use these columns along with SVM algorithms that you have learned to predict whether a purchase is fraudulent or legitimate.

a) Theoretically, based on the proportion of fraudulent purchases, determine the accuracy when considering all outputs as zero.
b) Considering part A and the dataset information provided to you, determine the importance of false positives and false negatives in evaluating the classifier.
c) Split the data into two sets—training and testing—using an 80:20 ratio.
d) Classify the training data using SVM (you can use other SVM types you've learned, such as SVM with kernels, in addition to linear SVM).
e) Evaluate the test data based on the classifier created and measure the accuracy of your classification. Pay particular attention to part B to ensure that the evaluation metrics used are practical and suitable.
f) What methods exist to increase accuracy in the invalid class? How does this affect the accuracy of each class? Does this method help reduce computations?

Problem 4: Diabet (20 points) - Romina

Load the “pima_indians_diabets.csv” dataset that is in the folder of exercise and classify the data using below models.

a) with at least 3 values for the below parameters, train the random forest classifier on the dataset and report the accuracy on the train and test dataset and specify the best model. (n_estimators, max_features, max_depth)

b) analyze the effects of the parameters on the model performance.

c) use any of the ensemble methods and try to achieve a better accuracy on the test dataset. (3 models is enough. But it’s better to achieve a better accuracy than the previous section model.)