Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Spring 2024

Teaching Assistants

Pooria Azizi (pooria.azizi@aut.ac.ir)

Amirmasoud Sepehrain (amirmasoud.sepehrian@aut.ac.ir)

Faezeh Hemmatzadeh (feazeh.hz@aut.ac.ir)

Nima Hatami (nima.h@aut.ac.ir)

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_03_[std-number].zip
- Report
  - ML_03_[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: Loan Approval Prediction | Decision Tree (20 + 10 points)

The given dataset in this question is the loan approval dataset which is a collection of financial records and associated information used to determine the eligibility of individuals or organizations for obtaining loans from a lending institution. The purpose is to train a decision tree model to predict loan approval based on an applicant's financial status. For this purpose, do the following steps (all the steps below should have a report):

a. Load the CSV file and show 10 random items, then visualize the data with at least 3 plots.
b. Perform the necessary data cleanings such as handling missing values, duplicates, outliers, incorrect values and etc.
c. Perform the necessary feature engineering, encoding and scaling.
d. Split the data into training and testing sets with 80/20 split ratio.
e. Initialize a decision tree classifier model with default parameters, fit the classifier to the training data and evaluate the performance of the model on the test data using metrics such as accuracy, precision, recall and f1-score.
f. Perform hyperparameter tuning to optimize the performance of the model. Use grid search to help you in this step. Then train and evaluate the model with best hyperparameters and compare the performance of this model with the previous one.
g. Visualize the decision tree using a library such as graphviz, scikit-learn or matplotlib. What information can be derived from this plot?
h. Perform pruning on the tuned decision tree model to prevent overfitting. Then train and evaluate it and compare it with the model before pruning. (Extra points)
i. Visualize this model and compare its plot with the previous one. (Extra points)

Problem 2: Poisonous Mushroom Classification | Logistic Regression (30 points)

In this question, we want to explore the application of logistic regression in classifying the edibility of mushrooms based on their physical characteristics.

a. Load the CSV file and show 10 random items. Display the output of the describe command for this dataset.

b. Perform the necessary data cleanings, feature engineering, encoding and scaling. The way that you do this part is very important. *(Tip: Use label encoding and one-hot encoding techniques for features properly)*

c. Split the data into training and testing sets with 70/30 split ratio.

d. Initialize a Logistic Regression classifier model, fit the classifier to the training data and evaluate the performance of the model on the test data using metrics such as accuracy, precision, recall and f1-score.

e. Run the logistic regression again, but this time using cross-validation, to ensure that you are not overfitting the data. A simple 10-fold cross validation is fine.

f. Compare the obtained results of (e) and (d).

Problem 3: Text Analysis | Naive Bayes (25 points)

The goal of this task is to determine the sentiment or emotion expressed in a piece of text, specifically using the IMDB movie review dataset. This dataset comprises 50,000 movie reviews labeled as either 'positive' or 'negative' sentiments. To train machine learning models for accurate classification of these reviews, follow these steps:

a. Load the CSV file of the IMDB dataset and display 10 random items.
b. Perform necessary preprocessing, including normalization techniques such as lemmatization, stemming, tokenization, and the removal of special characters.
c. Explain your approach to feature selection using both CountVectorizer and TfidfVectorizer models, employing two different strategies.
d. Split the IMDB dataset into training and testing sets, then perform data scaling.
e. Apply the Naive Bayes algorithm for sentiment classification. Evaluate the models using specified metrics and select the best-performing one. Finally, save the chosen model for future use.
f. Given a new review from the IMDB dataset, use the saved model to predict its sentiment. Provide the code for this prediction process.

Problem 4: Image Processing | Logistic Regression (25 points)

In this question, we intend to use the logistic regression model to recognize the handwritten numbers.

a. In the first step, please load the OpenCV digits image, divide it into its sub-images featuring handwritten digits from 0 to 9, and create their corresponding ground truth labels.
b. Split the data into training and testing sets with an 80-20 ratio.
c. Before training a classifier model using logistic regression, tell me how can convert a binary classifier into a multi-class classifier?
d. Train a logistic regression model as a multi-class classifier using your training and test datasets.
e. Report achieved accuracy. *(Note: Accuracy above 78% will be acceptable)*
f. As the final step, create a confusion matrix to gain a deeper understanding of which digits have been misidentified.

(*Note: If you need help solving this question, Google a little more as a first step)*