Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2024

Teaching Assistants

AmirHossein Babaeayan (Head TA)

Ali Raei

Assignment (1)

Outlines. In this assignment, we will focus on Preprocessing, Feature Engineering, and Data Cleaning.

Deadline. Please submit your answers before the end of October 12th 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 7 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:

ML_00_[std-number].zip
- Report
  - ML_0_[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. Good luck with your learning journey!

Preprocessing Job Experience Data for Predictive Modeling **by: Ali Raei**

In the era of digital transformation, job experience data has become crucial for making decisions related to talent acquisition, career progression, and human resource management. Accurately processing this data to extract meaningful insights is vital for machine learning models used in predicting career trajectories or job matching. Given a dataset, exp.csv, containing job titles, cooperation duration, company details, and locations, we aim to preprocess this data to make it suitable for machine learning tasks.
This assignment will guide you through various data preprocessing tasks necessary to clean, transform, and prepare the dataset for further predictive modeling applications. We will explore handling missing values, feature extraction, categorical encoding, and other essential preprocessing techniques.

Assignment Tasks:

a) Load the dataset exp.csv and show 5 samples.
   Explore the structure and initial contents of the dataset.
b) Handle Missing Data.
   Identify columns with missing values, and use appropriate imputation or removal techniques to clean the data.
c) Extract Numerical Features from Duration of Cooperation.
   The duration_of_cooperation column contains the start and end dates of job cooperation periods. Extract the total duration in months and create a new feature, cooperation_duration_months.
d) Text Preprocessing of Job Titles and Company Names.
   Clean and preprocess the title and company_name columns by removing unwanted characters, normalizing text, and identifying key job roles (e.g., "Manager," "Analyst," etc.).
e) Categorical Encoding of Location.
   Convert the location_hybrid column into numerical values using techniques like One-Hot Encoding or Label Encoding. Explain the choice of encoding and its impact on the model
f) Outlier Detection in Duration of Cooperation.
   Use z-score or IQR methods to detect outliers in the cooperation_duration_months column and remove them from the dataset. Discuss how outliers can affect predictive modeling.
g) Normalize Numerical Features.
   Normalize or standardize the cooperation_duration_months column to ensure uniformity in scale for modeling purposes. Discuss the impact of different normalization techniques on model performance.
h) Analyze Part-Time and Full-Time Job Experience.
   After converting the duration_of_cooperation data into timestamps, create visualizations to compare part-time and full-time job experiences. Plot the distributions of job durations for both categories and discuss any insights.
i) What is Model Preprocessing? How does it differ from Feature Engineering?
   Provide a brief explanation of how preprocessing prepares raw data for machine learning models and how it differs from feature engineering, which involves creating new variables or transforming existing ones.