Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfard

CE5501 | Fall 2023

Teaching Assistants

Zahra Akhlaghi

Zahra Zanjani

Atiyeh Moghadam

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 October 22th 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:

ML_01_[std-number].zip
- Report
  - ML_01_[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: Why and how (25 pts)

#Theorical

a) What are the requirements for a linear regression model to be accurate?
b) How do you deal with outliers in a linear regression model?
c) How many coefficients are required to calculate the relationship between a single predictor variable and a response variable in a univariate linear regression equation?
d) What is the difference between correlation and regression?
e) What is the difference between KNN regression and linear regression? Consider the following dataset. This dataset contains the length and weight of metal rods along with their cost. Calculate the cost for a rod with a length of 7 and a weight of 8 using k=3 and k=5. Write your calculations completely.

length	**weight**	**cost**
10	15	45
11	6	37
7	9	33
9	14	38
6	11	35
9	7	32
5	8	30
13	8	44

f) Fit a linear regression. Show only the first iteration of Gradient descent algorithm using learning rate of 0.02 for the following data, if the Relative Risk of Coronary Heart Disease is believed to be only linearly dependent on BMI as well as Diastolic Pressure. Assume the intercept of the regression model as 5 and the slope of independent variables as -0.03 (negative).

Patient	Systolic Pressure (mm Hg)	Diastolic Pressure (mm Hg)	BMI	Waist Circumference (cm)	RR-CHD (Relative Risk of Coronary Heart Disease)
1	140	80	35	100	1.81
2	120	80	25	80	1.22
3	130	100	30	60	1.71

Problem 2: Can I survive a surgery?

#Implementation

Fear of surgery has been a longstanding concern for many individuals over the decades. Whenever someone has faced the prospect of undergoing a surgical procedure, they have often questioned their chances of survival. In order to gain insights into this matter, let us delve into the available data. The dataset we are provided with pertains to the survival of patients who have undergone surgery for cancer. It contains information about age, year of surgery, number of positive axillary nodes and Survival status.

a) Why is it important to do exploratory data analysis?
b) Is there any missing values in dataset? How would you handle missing values?
c) Is the dataset balanced? How did you find out? Why is class imbalance problematic?
d) Report the average age of patients. Is the age of patients correlated with the chance of survival? Please explain.
e) Find the relationship between survival and the average number of auxiliary nodes.
f) Which of the three variables is more useful than the others in distinguishing between the classes 'yes' and 'no'? Support your claim with a suitable chart.
g) The presence of outlier data in a medical dataset is expected. Use a suitable chart to demonstrate the potential outliers in the dataset.
h) Finding the correlation among feature variables is particularly important when trying to determine the feature importance in regression analysis. Use a suitable chart to display the correlation between features. Then, utilize numerical evidence to support your observations. Do correlated features impact the performance of a statistical model?

Figure 1: According to cancer treatment and survivorship statistics . Survival rates have increased steadily since mid 1970s from 7% to 40% among adults aged 20 years and older.

Problem 3: Predicting Song Sales Using Machine Learning

#Implementation

Music plays a vital role in our digital age, with streaming and social media platforms fueling song sales. Various factors, such as musical features, social factors, and marketing strategies influence song sales. Machine learning offers a valuable tool for predicting song sales by analyzing features. This prediction capability can benefit music streaming services. The objective of this assignment is to utilize machine learning techniques to predict song sales based on musical features.

a) Load the dataset and identify the categorical and numerical features by counting the unique rows in each feature.
b) Analyze the distribution of the target variable and numeric features using suitable charts.
c) There are two main types of outlier detection: descriptive and prescriptive. Discuss the differences between these methods. Determine the suitable type of outlier detection for this problem and explain why. Use box plots to identify columns with outliers then remove the outliers from the dataset.
d) Investigate the relationships between all features using appropriate charts and provide detailed explanations.
e) Explain the concept of multicollinearity and discuss the appropriate techniques for handling it. Determine if multicollinearity is present in this dataset **(without using charts from part d)**. If multicollinearity exists, perform the necessary preprocessing steps.
f) **Implement linear regression from scratch** and compare the results and runtime with linear regression implemented in scikit-learn.
g) Discuss the difference between Lasso and Ridge regression. **Implement both models** and compare their performance with the previous linear regression results.
h) Explore the possibility of overfitting on the test or validation set. Discuss how you detect and avoid overfitting in this project.
i) Consider having intervals for the target value, where each interval corresponds to a specific category. In this case, we would have five different categories. We divided the target data into intervals of 10,000, where the range from 0 to 10,000 is labeled as the worst_seller, and the range from 40,000 to 50,000 is labeled as the best_seller. Discuss whether utilizing these intervals allows us to still use linear regression for prediction or if the problem is better suited for classification. Analyze the potential challenges associated with this approach.
j) Ensure model generalization by evaluating the model's performance using proper evaluation metrics for this problem. Analyze the errors and provide a detailed report.
k) Offer a new idea to improve the results of the discussed models (extra points).

Problem 4: Predicting House Price Using Stacked Regression

#Implementation

In today's dynamic real estate market, accurately predicting the sale price of residential properties is a critical challenge for both homeowners and real estate professionals. The sale price of a house is influenced by a myriad of factors, including location, size, condition, and various economic indicators. With the ever-changing landscape of housing markets, making informed decisions about buying or selling a property has become increasingly complex. To tackle this problem, we aim to harness the power of data science and machine learning techniques. In this endeavor, we will explore the development of a Stacked Regressions model that leverages the predictive capabilities of multiple regression algorithms.

a) Load the dataset and show 5 samples.
b) Use z-score to identify columns with outliers and then remove the outliers from the dataset.
c) Normalize the distribution of the target column
d) Preprocess the dataset.
e) Train Lasso, Elastic Net, Kernel Ridge, and Gradient Boosting Regression. Find the best parameters and learning rate in each model and explain in general how the learning rate affects the performance of the model (you can use sklearn library).
f) Report MSE and R2 on test data and compare all models.
g) What is model stacking? How does Stacked Regression work?
h) Train Stacked Regression model (You should implement Stacked Regression, and you can use sklearn library for base models)
i) Report MSE and R2 on test data and compare with the results in part f.