Amirkabir University of Technology
(Tehran Polytechnic)

Machine Learning Course By Dr. Nazerfadr

CE5501 | Fall 2022

Teaching Assistants

Mohsen Ebadpour(head) (M.Ebadpour@aut.ac.ir)

Ehsan Shobeiri (EhsanShobeiri@aut.ac.ir)

Mohamadreza Jafaei (Mr.Jafaei@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 October 28th 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_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 I (25 pts)

#Theorical

a) A regression task can be considered in three different topics: 1) as a least-square 2) as a linear programming and 3) as a convex optimization. Do the short research about them in regression task.

b) In a simple regression model with gradient descent, defining “iteration” and “complexity” is challengeable. Design a strategy to find suitable value for mentioned parameters.

c) We have a constant term(called θ_0) that it is not multiplied to any x in regression formula. Discuss about this term and explain what will be happen if we remove it.

d) Write a short comparison about K-fold and leave-one-out cross validation, then answer these questions:

   I. Which one is better for large-scale and small-scale datasets.

   II. Which one is robust against noisy dataset.

   III. Which one has lower computational complexity(time and memory).

   IV. Discuss and research about defining optimal value for K in K-fold cross validation.

e) The type of regularization that you talked about in class is popular for L-2 regularization. Study about L-1 regularization and explain Which one is better in different situations and why we need regularization terms in most of real-world tasks.

f) Assuming some noisy samples are added to the dataset in a regression task. Does it affect your ratio of dataset split to train/test and model complexity? Can you handle it with increasing iteration or model`s degree? Which one of gradient descent and normal equation has better result? Give supporting statements.

g) Assume a small dataset with blow table:

   I. Fit a linear regression with gradient descent and 3 iterations based on SSE loss function. Justify your answers with indicating all steps.

   II. Again, fit regression with normal equation.

X	0.9	0.5	0.3	0.7
Y	1.1	0.35	0.3	0.65

   III. Plot obtained regression lines; then report MAE, MSE and RMSE for previous parts and discuses about them.

h) In regression task with gradient descent, updating all learning weights for each sample in each iteration is a wrong strategy. Explain why, discuss about solutions and compare them [guide].

i) What is the difference/similarity between “regression” and “function estimation”?

Problem 2: warming up by implementing (15 pts)

#Implementation

Figure 1: In Iran, Health services and their relevant costs are a concern of the government`s employees, and they pay about 15% of their salary to mentioned titles.

Health is important but same as money? Based on high cost of health services, insurance companies founded and they take critical role for supporting costs. Analyzing the records may extract high-level information about behaviors or details. In this stage, you are given a small dataset of Artavil insurance company to analyze it and obtain some implementing experiences. To aim this, do the following parts:

a) Load the dataset, then plot a pie chart to indicate ratio of male/female among all customers.

b) We want to extract and report that people who have greater children have more charges at the same time or not. What kind of charts do you suggest? Explain why; then implement it and report the results.

c) Sub plotting is important to obtaining a few charts or figures in only one plot. Some mechanisms or libraries are implemented to aim that. One of them, is *subplot* in matplotlib of python. Using mentioned function or other methods that you are comfortable with, report four bar charts (2x2) in one plot to obtain combination of region, children and charges. Each plot will be related to one region, x-axis indicates children and y-axis shows mean charges of customers with same children and region.

d) Insurance company wants to know in what range of BMI, count of customers is greater than other ranges to offer some discounts. Plot a chart (what kind of chart!?) and extract a range with length of 1.5 to answer this goal. (Hint: You may plot a chart that shows density of *bmi* column)

e) Repeat previous part for age (length: 5) of customers and discuss about obtained ranges.

f) With desired chart(s), indicate that in what range of ages, smoker customers have lower charges against of non-smoker customers.

g) Now and according implemented parts, we want to determine effect of smoking on male/females based on charges and compare them. So, provide necessary charts in one plot and answer the questions below :

   I. Female smokers have more charges or male smokers?

   II. Male smokers have more charges or male non-smokers?

h) With a suitable chart(s), calculate in which regions smokers who have 3 children(at least) have more charges against of customer who are not smoker but have 4 children(at least)?

Note: Make sure that all details have been mentioned in all charts, figures, plots and reports. (titles, axises` label and etc.)

Problem 3: Math functions as regression (25 pts)

#Implementation

Math function is helping you to learn regression. Most of real-world problems have been solved by pure math functions or its properties. In this problem you will be faced to implement regression as well as math functions.

In image/signal processing topics there is artifact that is popular for aliasing as you see its sample[wiki]. It often occurs when sampling rate is too low.

A mathematical function in inverse Fourier transform of an image with the main lobe and some side lobes affects the image and an aliasing artifact appears (you will learn a lot about it in DIP course). You are given a dataset that consists of some samples by mentioned function; consider it and follow below parts:

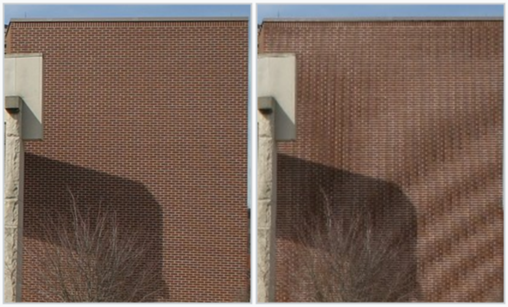

Figure 2: sample of aliasing artifact in image restoring.

a) Load and plot the dataset using a scatter plot.

b) Is dataset sorted? How do you check it and what is its time complexity? Do we always need shuffling the dataset in regression methods? If no, give an example. Explain your answers.

c) Split the dataset to train/test parts with the ratio of 8:2 and plot them together in one figure with different colors to indicate that splitting is done by balanced.

d) According to SSE loss function, implement below function based on gradient descent:

   RegressionTrainerGD (x, y, learning_rate, iteration, degree) : weights of regression model

   Implement your function such that a plot will be reported which indicates changes of MSE during the iterations.

   You are not allowed to use built-in functions or other tools like packages and etc. to aim train regression.

   Also, It is recommended to reduce learning rate during the iterations. (why?)

e) Using above function, train 15 regressions with combination of 100, 1000, 5000 iterations and degrees of 1,2,3,5,10. In one 3x5 plot, report obtained regression lines and in another 3x5 plot, report MSE for each train. Discuss about results and explain which one is optima.

f) Implement your strategy (1-b) and find mentioned parameter as well as reports of regression lines and MSE chart. Also, compare it with best result of previous part.

g) Now, implement normal equation method to train regression model (based on given formula in slides) and plot regression line then compare with results of previous parts and discuss about it.

h) Using the 5-fold cross validation with five different values for regularization term, try to train an optima regression model based on gradient descent with 2500 iterations and degree of 4. Report result and discuss about obtained value.

Note: mention all details in title of each chart.

Problem 4: Fish (20 pts)

#Implementation

Consider the Fish.csv dataset (target variable is “Weight”) then answer the following questions. For report accuracy model you should report R-Squared and R-Squared Prediction. you don’t need to split data. For all parts, you must explain the steps to solve the question:

a) Run linear regression and report the accuracy.
b) Run all-possible regression and report the accuracy table. Which one has the highest accuracy?
c) Run Stepwise method on this data set. Report the accuracy. Is the variable selected in stepwise different from the best result in the all-possible regression?
d) Does this dataset have multicollinearity problem? Show and explain.
e) Report DFFITS and Cook’s distance plot. Interpret these plots.
f) Analyze the residual error. Is it normal? If not, fix the problem and measure the accuracy again. Does normalizing the residual error fix the Heteroscedasticity problem?
g) Considering all the above, build your final model and report the accuracy. Accuracy above 99% has extra points. (Hint: you should use polynomial regression and variables interaction effect)

Problem 5: why and how II (15 pts)

#Theorical

a) What is Hat Matrix in Regression? Why is this matrix important?
b) What is the Stepwise in regression? Explain methods.
c) What is Multicollinearity and Heteroscedasticity? why it is a problem? How to detect and fix it?
d) What is DFFITS and Cook's distance? Explain.
e) Why do we assume that the residual error in the regression should be normal? If not, what should we do?