Assignment 1: Review of Linear Algebra and Basics of Regression

UVA CS 6316 :
Machine Learning (Fall 2018)

Out: W2
Due: 0915 Sun midnight 11:59pm @ Collab

a *The assignment should be submitted in the PDF format through Collob. If you prefer hand-writing QA parts of answers, please convert them (e.g., by scanning or using an app like Genuis Scan) into PDF form.*

b *For questions and clarifications, please post on Piazza.*

c *Policy on collaboration:*

Homework should be done individually: each student must hand in their own answers. It is acceptable, however, for students to collaborate in figuring out answers and helping each other solve the problems. We will be assuming that, with the honor code, you will be taking the responsibility to make sure you personally understand the solution to any work arising from such collaboration.

d *Policy on late homework:*

Homework is worth full credit at the midnight on the due date. Each student has three extension days to be used at his or her own discretion throughout the entire course. Your grades would be discounted by 10% per day when you use these 3 late days. You could use the 3 days in whatever combination you like. For example, all 3 days on 1 assignment (for a maximum grade of 70%) or 1 each day over 3 assignments (for a maximum grade of 90% on each). After you've used all 3 days, you cannot get credit for anything turned in late.

e *Policy on grading:*

***1**: 26 points in total, 2 points per question.*
***2**: 56 points in total. 10 points for code submission (and able to run), 6 point for successfully loading data, 40 points for correct implementation and good discussion of each optimization function (a total of 4).*
***3**: 18 points in total. 3 points for each sub question.*
The overall grade will be divided by 10 and inserted into the grade book. Therefore, you can earn 10 out of 10.

Please provide proper steps to show how you get the answers.

1 Linear Algebra Review (QA type)

Let $\mathbf{x} = (x_1, x_2, x_3)^T$ and:

$$\begin{cases} 2x_1 + 2x_2 + 3x_3 = 1 \\ x_1 - x_2 = -1 \\ -x_1 + 2x_2 + x_3 = 2 \end{cases} \tag{1}$$

Please answer the following questions:

- 1.1 Solve the linear equations
- 1.2 Write it into matrix form(i.e. $\mathbf{Ax} = \mathbf{b}$) (we will use the same $\mathbf{A}$ and $\mathbf{b}$ in the following questions.)
- 1.3 The Rank of $\mathbf{A}$ is ?
- 1.4 Calculate $\mathbf{A}^{-1}$ and $det(\mathbf{A})$
- 1.5 Use (1.4) to solve the linear equations
- 1.6 Calculate the inner product and outer product of $\mathbf{x}$ and $\mathbf{b}$.(i.e. $\langle \mathbf{x}, \mathbf{b} \rangle$ and $\mathbf{x} \otimes \mathbf{b}$)
- 1.7 Calculate the L_1,L_2 and L_∞ norm of $\mathbf{b}$
- 1.8 Now we add one more linear equation $-x_1 + 2x_2 + x_3 = 1$ into linear equations above. Write it into matrix form(i.e. $\mathbf{A_1 x} = \mathbf{b}$)
- 1.9 The rank of $\mathbf{A_1}$ is?
- 1.10 Could these linear equations be solved? Why ?
- 1.11 Calculate the Pseudo-inverse of $\mathbf{A_1}$. (You can use a tool to solve it.)
- 1.12 Suppose $\mathbf{B} = \begin{bmatrix} -\frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \\ \frac{2}{3} & \frac{2}{3} & \frac{1}{3} \end{bmatrix}$
  Is $\mathbf{B}$ an orthogonal matrix? Why ?
- 1.13 Suppose $\mathbf{y} = (y_1, y_2, y_3)^T$, calculate $\bigtriangledown_{\mathbf{y}} \mathbf{y}^T \mathbf{A} \mathbf{y} = \frac{\partial (\mathbf{y}^T \mathbf{A} \mathbf{y})}{\partial \mathbf{y}} =$;

2 Linear Regression Model Fitting (Programming)

There are **TWO** portions to this section.

2.1 Coding

First, you must fill out the template code provided. This code will perform linear regression on a data file named "regression-data.txt" which is also provided.

(BTW: This coding assignment does not include validation or testing which is bad practice but we will do them later in the semester.)

Please submit your python code as "linearRegression.py" and use Python3.

Using the "numpy" arrays and functions from `http://www.numpy.org` is required. (BTW: Numpy arrays and functions perform mathmatical operations faster because they allow for vectorization and use optimized libraries. For this dataset size, it is irrelevent, but for larger datasets it means the difference between waiting 1 hour and 4 days. So we are forcing you to use them as practice.

2.2 Written Portion

Second, you must complete a written portion and turn it in as part of the pdf with the rest of the assignment. For the written portion you must describe what happens to the loss function per epoch as the learning rate changes for both gradient descent and stochastic gradient descent and explain why.

- Function load_data_set() should also output a figure plotting the data; Please submit the plot in the written part of the homework.
- For each optimization method you implement to learn the best LR line, please submit a figure showing the data samples and also draw the best-fit line which has been just learned. Please also include the concrete value of the derived theta in the written part of the homework.
- For each optimization method you implement to learn the best LR line,
- In your GD, SGD or MiniSGD implementation:
  - The functions should output a figure with x-axis showing epoch number (the updating iteration t, t mean the total number of times the entire training set is iterated over), and y-axis showing the mean training loss at that epoch.
  - The functions should use mean square error(MSE) as the loss function.
  - The functions should stop when t reaches a predefined value $t_{max} = 100$;
  - In the written part of the submission, you should discuss the behavior of the function when varying the value of the learning rate (for instance varying values from $\{0.001, 0.005, 0.01, 0.05, 0.1, 0.3\}$.
  - It is a good practice to perform a random shuffle your training samples before each epoch of (mini-batch) SGD;
  - In mini-batch SGD, you should try to observe the convergence behaviors by varying the size B your used for sizing the mini-batch.

2.3 Recommendations

- Implement the code in order that it is given in the template main function.
- Shuffle the x and y before using stochastic gradient descent. (Make sure to shuffle them together.)
- 0s will not work as hyperparameters for learning rate or number of iterations.

We will run "python3 linearRegression.py" and it should work!
A few useful links for using numpy array:

- basics (more than we need): `https://docs.scipy.org/doc/numpy/user/quickstart.html`

- a good comprehensive list of math operations we normally need to use on numpy arrays `https://www.numpy.org/devdocs/user/numpy-for-matlab-users.html#array-or-matrix-which-should-i-use`

3 Sample Exam Questions:

Question 3.1. Linear Regression+ Train-Test Split

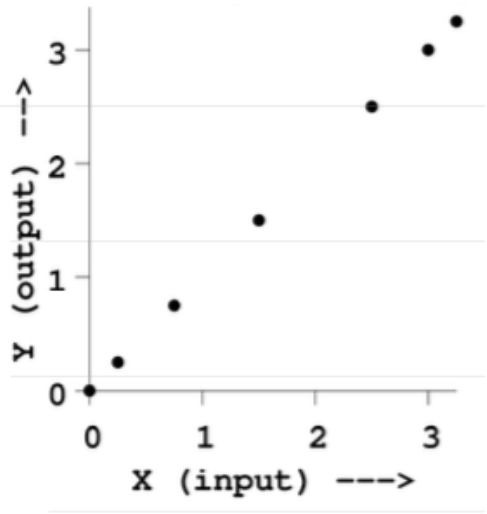


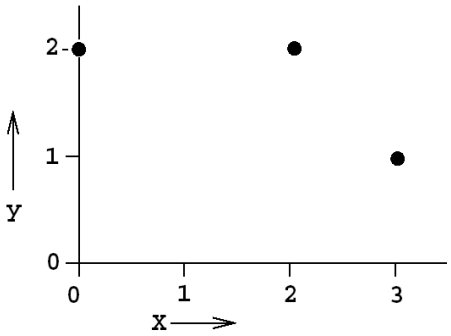
Figure 1: A reference dataset for regression with one real-valued input (x as horizontal axis) and one real-valued output (y as vertical axis).

What is the mean squared training error when running linear regression to fit the data ? (i.e., the model is $y = \beta_0 + \beta_1 x$). Assuming the rightmost three points are in the test set, and the others are in the training set. (you can eyeball the answers.)

Question 3.2. Linear Regression+LOOCV

Note: LOOCV means (Leave-One-Out-Cross-Validation)

Suppose you are given a data set with three data points (see both table and figure). This dataset includes one real-valued input variable and one real-valued output variable. We are trying to learn a regression function to map from the input to the output. (Key: mostly likely you only need 10 minutes to solve the following three sub-questions.)



x	y
0	2
2	2
3	1

(you can eyeball the answers.)

(a) What is the mean squared leave one out cross validation error if using linear regression ? (i.e. the mode is $y = \beta_0 + \beta_1 x$)

(b) Suppose we use a trivial algorithm of predicting a constant $y = \beta_0$ (Att: you can assume any function form to map from input to output ! This is your assumption !). What is the mean squared leave one out error in this case? (Assume β_0 is learned from the non-left-out data points.)

(c) Based on the LOOCV results, which model category will you pick for this dataset: the category described in (a) or (b)?

Question 3.3. Error Evaluations for Regression

We are given a dataset with R records in which the i^{th} record has one real-valued input attribute x_i and one real-valued output attribute y_i. By running a linear regression method on this data, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set. Now let us increase the training set size gradually. As the training set size increases, what do you expect will happen with the mean training and mean testing errors? (Please provide 1 sentence of explanation to justify your choice.)

(e) Mean Training Error:

- A. Increase;
- B. Decrease

Your choice:

(f) Mean Testing Error:

- A. Increase;
- B. Decrease

Your choice: