

Assignment 2: Linear and Ridge Regression

UVA CS 6316/4501 :
Machine Learning (Fall 2016)

Out: Sept. 15 / Thu, 2016
Due: Sept. 25 / Sun midnight 11:55pm, 2016 @ Collab

- a** *The assignment should be submitted in the PDF format through Collob. If you prefer hand-writing the writing part of answers, please convert them (e.g., by scanning or using PhoneApps like officeLens) into PDF form.*
- b** *For questions and clarifications, please post on piazza. TA Jack (jkl5sw@virginia.edu) or Muthu (mc4xf@virginia.edu) will try to answer there.*
- 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 15% 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 55%) or 1 each day over 3 assignments (for a maximum grade of 85% on each). After you've used all 3 days, you cannot get credit for anything turned in late.*

Question 1. Linear Regression Model Fitting (TA: Jack)

A data file named “Q2data.txt” is provided. For this data, you are expected to find best-fit lines with linear regression model. Please submit your python code as “linearRegression.py” and follow the following required function names. Using the “numpy” package from <http://www.numpy.org> is recommended.

- The first function is to open a text file “Q2data.txt” with tab-delimited values and assumes the last value is the target value. This function should also output a figure plotting the data; Please submit the plot in the written part of the homework.

```
xVal,yVal=loadDataSet('Q2data.txt')
```

- The second function is to find the best-fit line with a linear regression model for the provided dataset. Solving this problem is one of the most common applications of statistics, and there are a number of ways to do it other than the normal equation based method.

You can choose one of the optimization methods we learned (normal equation, gradient descent, stochastic gradient descent) in the class. This function should output a figure showing the data samples from “Q2data.txt” and also draw the best-fit line which has been just learned. Please present the derived theta and the plot in the written part of the homework.

```
theta = standRegres(xVal,yVal)
```

- The third function is to find the best-fit line with a *polynomial* regression model for the provided dataset. As discussed in class, there are many datasets where a standard linear regression model is not sufficient to fit the data (i.e. the data is not linear). Thus, we need a higher order model to more accurately fit the data. For this function, you must use at least a 2nd order polynomial regression model for the provided dataset (if you want to use a higher order than 2, you may).

As with the previous function, you can choose one of the optimization methods we learned during the class. Recall from the lecture 5 notes that the problem of learning the parameters is still linear although we are learning a nonlinear function. This function should output a figure showing the data samples from “Q2data.txt” and also draw the best-fit line which has been just learned. Please present the derived theta and the plot in the written part of the homework.

```
theta = polyRegres(xVal,yVal)
```

We should be able to run “python linearRegression.py” and it should work!

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

ATT: 0.5 extra credit will be given to students who correctly implement TWO different optimization strategies and present/discuss these results correspondingly.

Question 2. Ridge Regression (TA: Muthu)

- Purpose 1: To emphasize the importance of selecting the right model through k-folds CV when using supervised regression.
- Purpose 2: To show a real case in which linear regression learns badly and adding regularization is necessary.


This problem provides a case study in which just using a linear regression model for data fitting is not enough. Adding regularization like ridge estimator is necessary for certain cases.

- Here we assume $X_{n \times p}$ represents a data sample matrix which has p features and n samples. $Y_{n \times 1}$ includes target variable's value of n samples. We use β to represent the coefficient. (Just a different notation. We had used θ for representing coefficient before.)
- 1.1 Please provide the math derivation procedure for ridge regression (shown in Figure)

Figure 1: Ridge Regression / Solution Derivation / 1.1

- If not invertible, a solution is to add a small element to diagonal

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p \quad \text{Basic Model,}$$

$$\beta^* = (X^T X + \lambda I)^{-1} X^T \bar{y}$$
- The ridge estimator is solution from 

$$\hat{\beta}^{ridge} = \operatorname{argmin} (y - X\beta)^T (y - X\beta) + \lambda \beta^T \beta$$

(Hint1: provide a procedure similar to how linear regression gets the normal equation through minimizing its loss function.)

(Hint2: $\lambda \|\beta\|_2 = \lambda \beta^T \beta = \lambda \beta^T I \beta = \beta^T (\lambda I) \beta$)

(Hint3: Linear Algebra Handout Page 24, first two equations after the line “To recap,”)

- 1.2 Suppose $X = \begin{bmatrix} 1 & 2 \\ 3 & 6 \\ 5 & 10 \end{bmatrix}$ and $Y = [1, 2, 3]^T$, could this problem be solved through linear regression?

Please provide your reasons.

(Hint: just use the normal equation to explain)

- 1.3 If you have the prior knowledge that the coefficient β should be **sparse**, which regularized linear regression method should be chosen to use ? (Hint: sparse vector)
- A data file named “RRdata.txt” is provided. For this data, you are expected to write programs to compare between linear regression and ridge regression.
- Please submit your python code as “ridgeRegression.py” . Please use the following instructions and use required function names. Please use Numpy or other related package to implement the ridge regression. Other requirements or recommendations are the same as Homework1.
- Notation: The format of each row in data file is $[1, x_1, x_2, y]$, where x_1, x_2 are two features and y is the target value.
- 1.4 For “ridgeRegression.py”,

- Load the data file and assume the last column is the target value. You should use $xVal$ to represent the data sample matrix and $yVal$ to represent the target value vector.
- 1.4.1 The first function is to implement the ridge regression and return the coefficient β with the hyperparameter $\lambda = 0$. (i.e. when $\lambda = 0$, it's just the standard linear regression). Please plot the data points and the learned plane ¹. Please submit the result into the writing part of this assignment. You are required to provide the following function (and module) for grading:

$$betaLR = ridgeRegression.ridgeRegress(xVal, yVal, lambda = 0)$$

- 1.4.2 The second function is to find the best λ by using a 10-fold cross validation procedure. The function should be,

$$lambdaBest = ridgeRegression.cv(xVal, yVal)$$

- You don't need to regularize the β_0 . Instead, you can estimate β_0 by center the input(i.e. $\hat{\beta}_0 = \frac{\sum y_i}{n}$).
 - (Hint1: you should implement a function to split the data into ten folds; then loop over the folds; use one as test, the rest train)
 - (Hint2: for each fold, on the train part, perform ridgeRegress to learn β_k ; Then use this β_k on all samples in the test fold to get predicted \hat{y} ; Then calculate the error (difference) between true y and \hat{y} , sum over all testing points in the current fold k .)
 - Try all the λ values from the set: $\{0.02, 0.04, 0.06, \dots, 1\}$ (i.e. $\{0.02i | i \in 1, 2, \dots, 50\}$). Pick the λ achieving the best objective criterion from the 10-fold cross validation procedure. Our objective criterion is just the value of the loss function (i.e. $J(\theta)$ MSE in the slides) on each test fold. Please plot the λ versus $J(\beta)$ graph (which is also called path of finding the best λ) and provide it into the writing.
 - Note : To constrain the randomness, please set seed to be 37. ²
 - Then run the ridge regression again by using the best λ calculated from 1.4.2. Please include the result into writing.
- $$betaRR = ridgeRegression.ridgeRegress(xVal, yVal, lambdaBest)$$
- Please plot the data points and the learned plane from best ridge regression. Please include the result into writing. ³.
 - 1.5 If assuming the true coefficient in problem 1.4 is $\beta = (3, 1, 1)^T$, could you compare and conclude whether linear regression or ridge regression performs better ? Explain why this happens based on the data we give.
 - (Hint: 1. Please implement a standard linear regression between x_1, x_2 and plot the x_1 versus x_2 graph;)
 - (Hint: 2. Guess the relationship between the two features and consider the problem 1.2.)

¹http://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html#surface-plots

²More about random in python, please see, <https://docs.python.org/2/library/random.html>

³http://matplotlib.org/mpl_toolkits/mplot3d/tutorial.html#surface-plots