University of Colorado at Colorado Springs

Home Work Assignment 1: Regression

Out: 02-10-2021, Due: 03-01-2021

Preface

In this homework assignment, you will answer questions as asked and write programs as necessary. You will also submit a nicely written report. You can do more if you want for extra credit. Consider the homework as open-ended and as a result, you may extend it any way you can. Extra credit will be given for substantial additional work. Please read on your own if a topic has not been discussed in class to your satisfaction. *Make sure you have a demo scheduled with me, the week the homework is due or the week after.* Note that you will have to keep working on your semester project as you work on this and future homeworks; so please manage time properly.

Introduction

We will explore how several types of regression work. As we all know, regression requires us to solve an optimization problem. Different types of regression set up this problem differently.

We have also discussed some simple algorithms for solving such optimization problems. Least Squares Linear Regression (LSLR), as discussed in class, solves the optimization problem quite simply. However, in general we have to use a method like *Newton-Raphson*, *Gradient Descent*, or variations of these to solve optimization problems.

Assume that we are given a dataset D, which is composed of a N examples. Each training row has a description of the example in terms of n features, and a target value or label y. Thus, the ith training example is given as

$$\langle x_1^{(i)}, \cdots x_n^{(i)};\ y^{(i)} \rangle.$$

Visually, the training dataset can be seen like the matrix or table as given below.

No	$x_1 \cdots x_n$	y
1	$\cdots$	.
2	$\cdots$	.
$\vdots$	$\vdots$	$\vdots$
i	$x_1^{(i)} \cdots x_n^{(i)}$	$y^{(i)}$
$\vdots$	$\vdots$	$\vdots$
N	$\cdots$	.

Least Squares Linear Regression and Beyond

For linear regression we assume that the target is numeric. The set of hypotheses in LSLR is simply the set of all straight lines. This is also called the Inductive Bias of the algorithm in that the algorithm knows from the beginning that it is trying a fit a straight line. In particular, we fit a linear function

$$h_{\vec{\theta}}(x) = \theta_0 + \theta_1 x$$

to the dataset if the training examples are scalar, i.e., each example has one feature x. If a training example is a vector of n features, the function we fit is

$$h_{\vec{\theta}}(x) = \vec{\theta}^T \vec{x} = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

where $\vec{\theta}^T = [\theta_0 \cdots \theta_n]$. This is the equation of a plane in n dimensions. $\vec{\theta}$ is a vector of parameters. Assume all feature values are numeric. Consider all vectors to be column vectors, although they may be written out as a sequence or in some other manner.

There are many other forms of regression, which may be able to solve a regression problem better than LSLR. You can find detailed discussions in Elements of Statistical Learning [1]. Three methods we are particularly interested in are Ridge, Lasso and Elastic Net.

R and Python have libraries that implement these regression methods.

To Do

Please do the following. Since you are going to submit a nicely written lab report (paper), you will type the material so that it is easier to read.

1. State the objective function that needs to be solved for Least Squares Linear Regression.
2. There are many ways to solve an objective function like this one. We discussed the standard calculus-based approach for it in class. In practice, the direct calculus-based approach is not generally used. An approach commonly used for solving optimization problems in machine learning is stochastic gradient descent (SGD).

   Implement SGD from scratch to solve the LSLR problem, using a programming language of choice. You can use math libraries, but not an existing implementation of SGD. You can use any language you want. However, you will have to demo it to me.

   Discuss SGD briefly in your paper. Discuss your implementation.
3. Test your implementation on the *Combined Cycle Power Plant* dataset from the UCI Machine Learning Repository. In addition, find another dataset in the UCI repository to test.

   Report your results. In other words, give the equation of the fitted line. Compute one or more measures of goodness of fit; these are also called evaluation metrics. Briefly discuss the metric(s) you use.

4. Write a program to implement Ridge regression using SGD. You should be able to use the SGD routine you wrote earlier. Change the hyperparameter value for Ridge regression to see how the results change. Solve the Combined Cycle Power Plant problem using Ridge regression. Use also the second dataset you chose earlier.

   Write the objective function to optimize for Ridge regression. Discuss the results you get using Ridge regression as you change the hyperparameter's value. Discuss any problems you face in all your implementations and how you take care of them.
5. Learn how to solve problems with LSRL, Ridge, Lasso and Elastic Net Regressions, either in R or Python, using existing libraries. Compare the equations that you obtain with the libraries with what you got earlier with your own implementations.

What to Hand in

You will submit a 2-4 page paper with a title and your name. Use the AAAI Author style in LaTeX you have been using for the semester project papers you have been writing for the class. In the paper, you will have a short section with an appropriate heading for each question asked and any extra work you perform. You must have a section called *Results*, in addition to other sections you deem appropriate. Here, you will present your findings in terms of discussions, tables, graphs and any other visuals as yo!u deem appropriate. It is always a good idea to start each paper with an *Introduction* section and close it with a *Conclusions* section.

Conclusions

As mentioned earlier, please do what you have been tasked with above. Perform additional research, do additional implementations, perform additional experiments, etc., if you want Extra Credit.

References

[1] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1, no. 10. New York: Springer series in statistics, 2001.