University of Colorado at Colorado Springs

Home Work Assignment 1: Regression

Out: 02-09-2018, Due: 03-05-2018

1 Preface

In this homework assignment, you will answer questions as asked and write programs as necessary. You can do more if you want for extra credit. Consider the homework is open-ended and as a result, you should do the basics and extend it any way you can. Extra credits 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 the TA, Marc Moreno Lopez, the week the homework is due.* Please 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.

2 Introduction

We will explore how several types of regression work. We have discussed several regression algorithms in class. As we all know, regression requires us to solve an optimization problem. Different types of regression do so differently.

We have also discussed (or will discuss) 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* or *Gradient Ascent/Descent* or variations of these to solve complex optimization problems.

Assume that we are given a dataset D, which is composed of a m 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

$$< x_1^{(i)}, \cdots x_n^{(i)};\ y^{(i)} > .$$

Visually, the training dataset can be seen like the matrix/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	$\cdots$	.

3 Least Squares Linear Regression

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

$$h_\theta = \theta_0 + \theta_1 x$$

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

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

where $\theta^T = [\theta_0 \cdots \theta_n]$. This is the equation of a plane in n dimensions. θ 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.

3.1 To Do

Please do the following. Since you are going to submit a nicely written lab report (paper), you should 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. Starting from the objective function, derive the equation of a straight line that optimizes it. Show steps.
3. Write a program to find the equation of the LSLR line given a training dataset. You can use any language you want. However, you will have to demo it to Marc.
4. Test it on the "Combined Cycle Power Plant" dataset from the UCI Machine Learning Repository.
5. Report your results, i.e., give the equation of the fitted line. In addition, compute one or more measures of goodness of fit; these are also called evaluation metrics.

4 Logistic Regression

Consider that we have a dataset where an example's features are all numeric as before, but the dependent variable or the target value or the label is either 0 or 1. We fit the sigmoid or logistic function $\sigma(z) = \frac{1}{1+e^{-z}}$ to the data so that we can actually build a binary classifier. In our case, z will be a linear function of x, the feature vector.

Assume that the ith training example $x^{(i)} =< x_1^{(i)}, \cdots x_n^{(i)}; y^{(i)} >$ where $y^{(i)}$ is either 0 or 1.

4.1 To Do

1. Given a dataset for binary classification using logistic regression, derive the optimization function.

2. Given this optimization function, you can optimize it in many ways. Two simple approaches to optimization are (a) Newton-Raphson Method and (b) Gradient Ascent/Descent Method. Write and briefly describe the algorithms for these two approaches, and how they are applied to solve the logistic regression problem.

3. Implement one of the two methods, using any language of your choice. You can use numeric libraries, but no direct implementations of the two methods.

4. Test it on the Iris and Wisconsin Breast Cancer datasets from the UCI Machine Learning Repository and report your results. In reporting your results, write out the fitted equation as well as some measures of goodness of fit.

5 Playing with Tools/Libraries

Even though this is listed as the last problem to work on, it may be beneficial if you start with this, so that you understand how "professional" tools implement regression.

5.1 To Do

1. Install Weka and R on a computer you can bring for demo to school. Learn how to perform LSLR and Logistic Regression using these two tools. Demo running them to Marc for the datasets mentioned earlier.

6 What to Hand in

You will submit a 2-4 page paper with a title and your name. Use the AAAI Author style 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. Describe 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.

I'll ask Marc Moreno to schedule demos the week the homework is due.

7 Conclusions

As mentioned in the onset, please do what you have been tasked with above. Perform additional research, do additional implementations, perform additional experiments, etc., if you want Extra Credit. Be enthused about Regression and Machine Learning!