September 21, 2017 Lecture 4, Sep. 18, 2017

Linear Regression

Lecturer: Justin Domke *Scribe: Lakshmi Vikraman*

1 Summary

Last class we talked about KNN for regression and its pros and cons. Today we will cover Linear Regression, including examples to understand it intuitively as well as mathematical derivations to understand it theoretically. Apart from these, a few administrative details would also be covered.

2 Linear Regression

2.1 Model

Let $X = [X_1, X_2, ..., X_p]$ be a input vector in $\mathbf{R}^p$ dimensional space. We refer to the p entries as either inputs, features, attributes, predictors. Let Y be the output variable which is a scalar in regression *i.e.* $Y \in \mathbf{R}$.

We can then define a function

$$f(X) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

where $\beta = [\beta_0, \beta_1, \beta_2, ...\beta_p]$ are the unknown parameters of the model which can be represented as a vector in $\mathbf{R}^{p+1}$ dimensional space. β_0 is used to shift the curve up or down and it can be folded into the summation part of the equation by adding a constant of 1 to X i.e. $X_0 = 1$, which makes X a $p+1$ dimensional vector. The summation then becomes an inner product of β and X and hence the function can be written as

$$f(X) = \beta^T X$$

2.2 Parameter Estimation

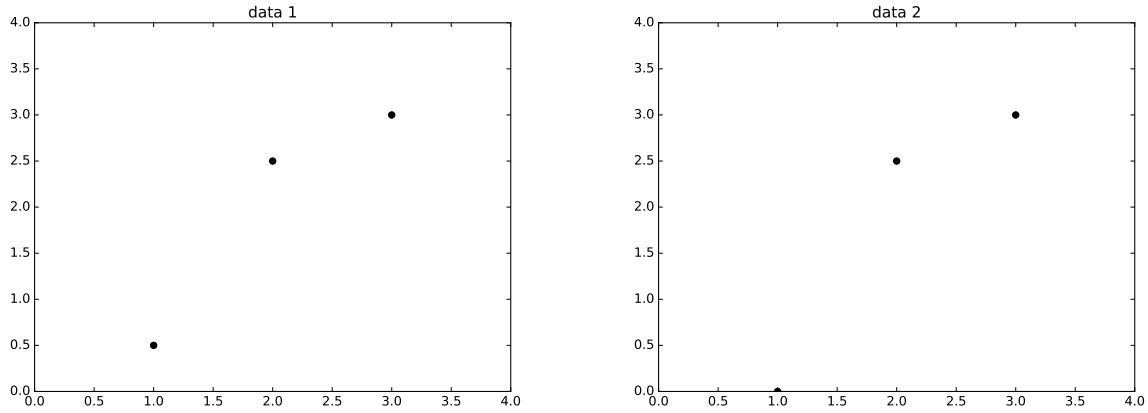
Given a dataset D, of X, Y pairs of the form $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ of n examples where each x_i is a $p+1$ dimensional vector, how do we estimate β such that $y_i \approx f(x_i)$.
The stategy is to provide a quantitative measure of "how well" a given β fits to a dataset (which is called a loss function) and then find the β which best optimizes this function. Two methods to measure this quality of fit is Mean Squared Error (MSE) and Mean Absolute Error(MAE) which is defined below.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i)]^2$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|$$

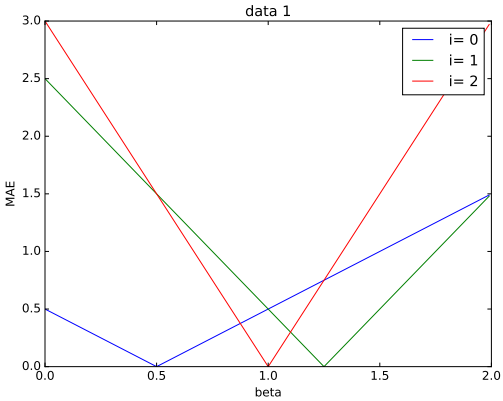
2.2.1 Examples

In order to understand what it means to find the best β , let us look at two toy datasets data1 and data2 shown below in figures 2.1a) and 2.1b) where p=1 ie X is one-dimensional. The datasets data1 and data2 only differ in their y_i at datapoint $i = 0$. The graphs illustrate how the value of the β varies with respect to the two loss functions. The figures 2.2a , 2.2b ,2.3a and 2.3b shows how the MAE of individual datapoints(indicated by blue,green and red curves) and mean value(black curve) varies with β for each of the two datasets. Similarly the figures 2.4a , 2.4b ,2.5a and 2.5b shows the behaviour for MSE . On comparing the two figures 2.2aand 2.3a for data1 and data2 respectively, we note that the left most "V" shifts to the left and β^* remains the same.Similarly on comparing figures 2.4b and 2.5b for MSE, we can see that left curve shifts slightly and β^* decreases slightly.

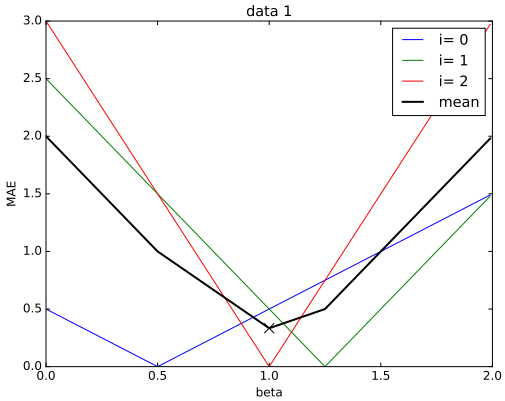


(a) Data1 (b) Data2

Figure 2.1

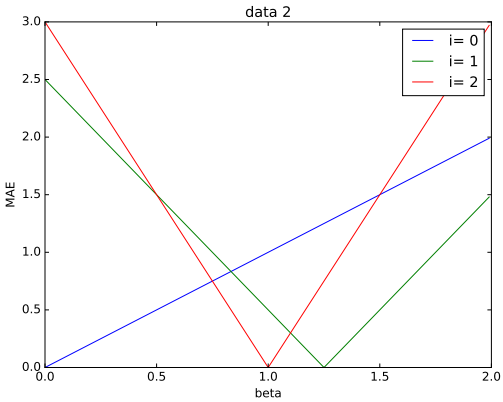


(a) MAE vs β

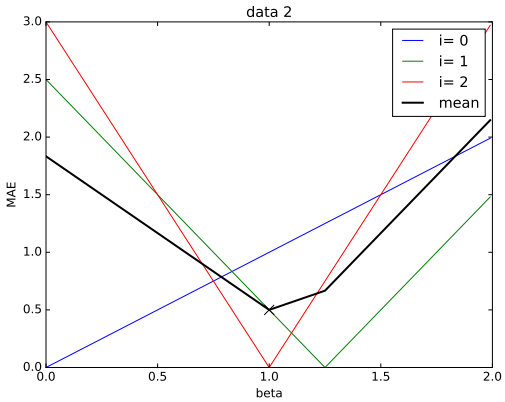


(b) MAE mean vs β

Figure 2.2

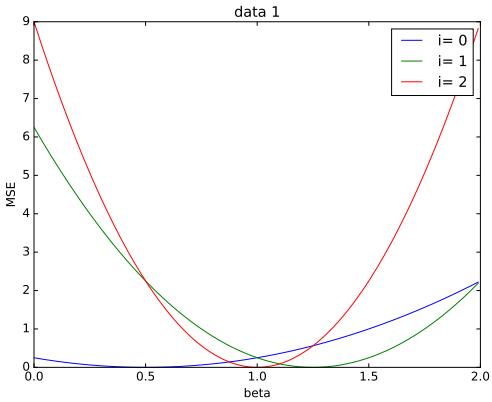


(a) MAE vs β

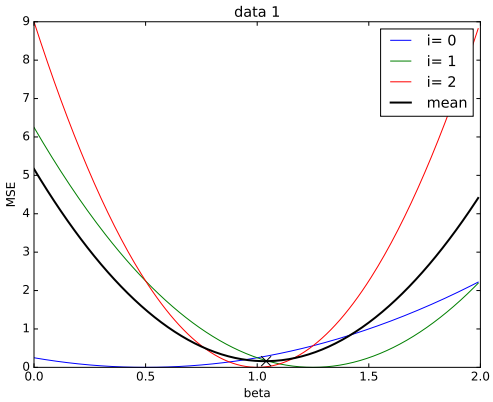


(b) MAE mean vs β

Figure 2.3

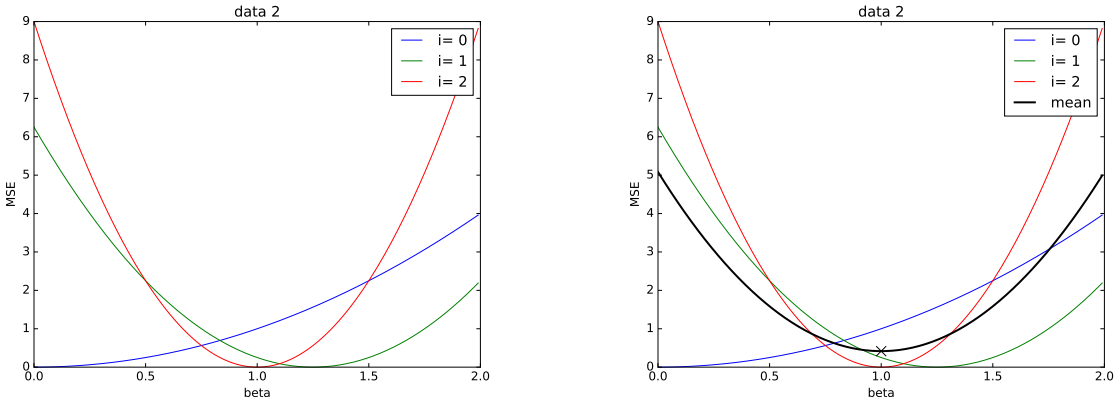


(a) MSE vs β



(b) MSE mean vs β

Figure 2.4



(a) MSE vs β (b) MSE mean vs β

Figure 2.5

2.2.2 Ordinary Least Square for p=1

Now let us the derive the parameter β for the case where p=1 . This method is called Ordinary Least Square(OLS) and finds the optimal β (β^*) which minimizes the MSE on the training set. *i.e.*

$$\text{Minimize } RSS(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

$$\text{Since for } \beta^* \ \frac{RSS(\beta)}{d\beta} = 0$$

$$\frac{RSS(\beta)}{d\beta} = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\beta}(y_i - \beta x_i)^2$$

$$\frac{RSS(\beta)}{d\beta} = \frac{1}{n}\sum_{i=1}^{n}2(y_i - \beta x_i)(-x_i)$$

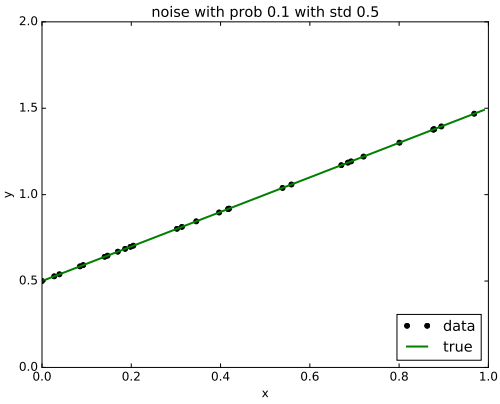
$$\frac{RSS(\beta)}{d\beta} = \frac{2}{n}\sum_{i=1}^{n}\beta {x_i}^2 - \frac{2}{n}\sum_{i=1}^{n}x_i y_i = 0$$

$$\beta^* = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}$$

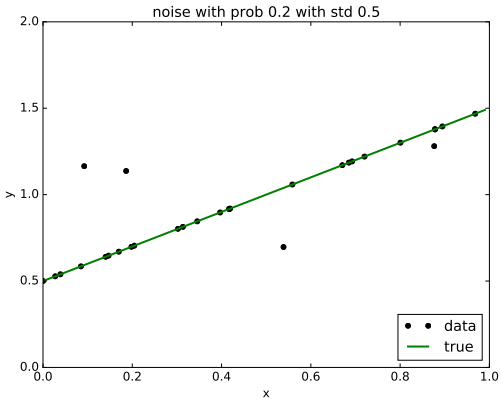
2.2.3 More Examples

Let us look at a few more examples where data $X_i \in [0,1]$ and is generated randomly. This data is used to fit the "true" curve $Y_i = 0.5 + X_i$ which is represented by green lines in figures

2.6a , 2.6b , 2.7a and 2.7b. With some probability we add gaussian noise with some standard deviation to this data X and then use this to fit the curve. The noisy data is represented as dots in the Figures 2.6a , 2.6b , 2.7a and 2.7b. The MSE and MAE is plotted against the data and shown in figures 2.8a , 2.8b , 2.9a and 2.9b.

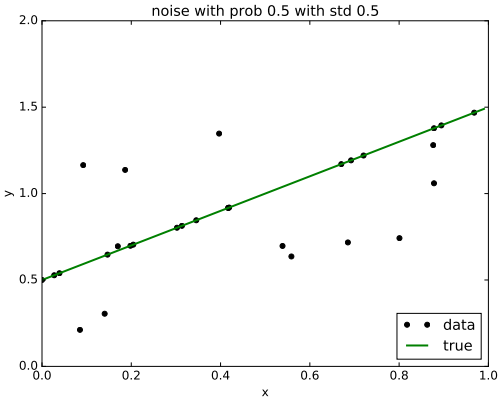


(a) Prob 0.1 and SD 0.5

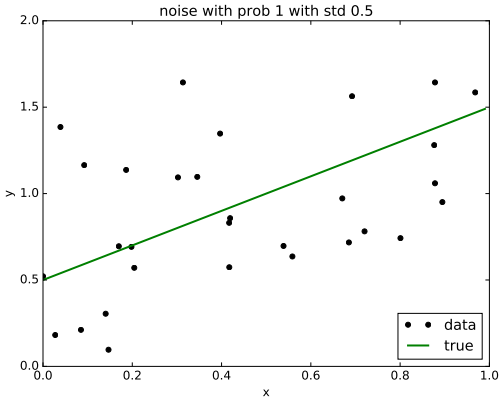


(b) Prob 0.2 and SD 0.5

Figure 2.6

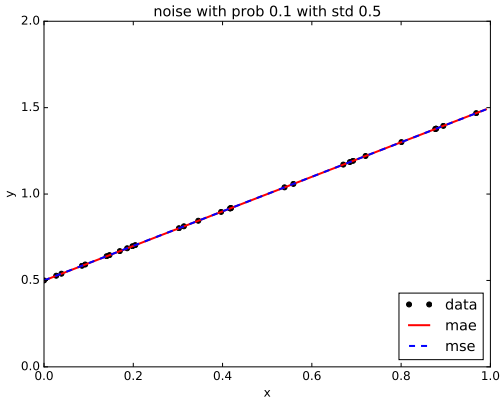


(a) Prob 0.5 and SD 0.5

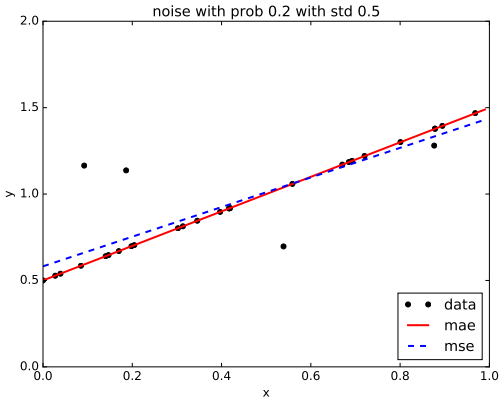


(b) Prob 1.0 and SD 0.5

Figure 2.7

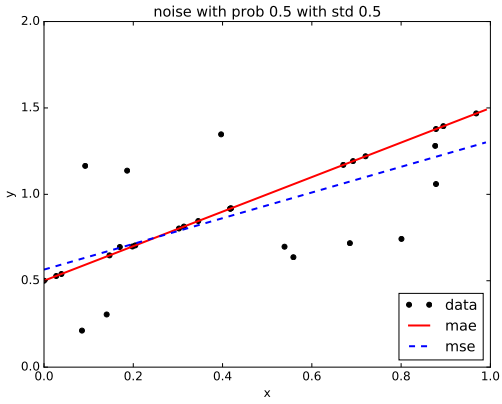


(a) Loss for noise with Prob 0.1 and SD 0.5

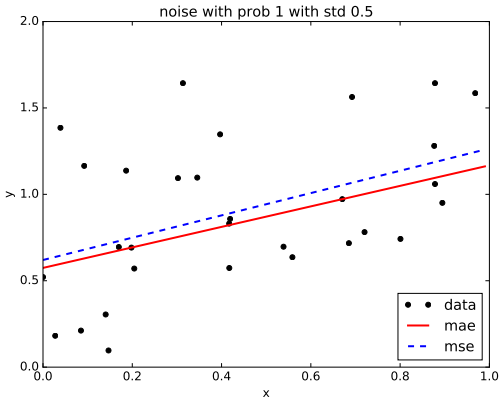


(b) Loss for noise with Prob 0.2 and SD 0.5

Figure 2.8



(a) Loss for noise with Prob 0.5 and SD 0.5



(b) Loss for noise with Prob 1.0 and SD 0.5

Figure 2.9

2.2.4 Discussion

Based on the examples , we can note that in case of MSE, each datum contributes a smooth loss function and each data point impact the result for optimal β^* while in case of MAE, each datum contributes a "V" shaped curve and changing some of the data may not impact the result for optimal β^*. In general, which loss function must be used for a task is "user specified" and there is no "right answer" for this.

2.2.5 General OLS for input vector X with p dimensions

Lets derive Ordinary Least Square solution for case where X is a vector in $\mathbf{R}^p$ dimensional space *i.e.*

$$RSS(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta x_i)^2$$

and we have to find β which minimizes $RSS(\beta)$.

Let us introduce some notation. Given a dataset D, of X, Y pairs of the form $D = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$ of n examples let $\mathbf{Y}$ represent a n dimensional vector where each dimension corresponds to y_i. Let $\mathbf{X}$ represent a n by p matrix where each row corresponds to x_i. Hence RSS can be written as follows.

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta)$$

$$\text{At minimum}, \nabla_\beta RSS(\beta) = 0$$

$$\nabla_\beta RSS(\beta) = \begin{bmatrix} \frac{\partial RSS}{\partial \beta_1} \\ \frac{\partial RSS}{\partial \beta_2} \\ \vdots \\ \frac{\partial RSS}{\partial \beta_p} \end{bmatrix}$$

This is a system of p equations with p unknowns. We will continue with the derivation in the next lecture.

3 Few Administrative Details

- You should have access to the following resources : Gradescope, Piazza and Moodle. If you face any difficulties in accessing them, please mail the head TA at anataraj@cs.umass.edu. You can also modify the associated email addresses if needed.
- The Instructor's office hours would be held on Monday from 3:45 - 4:45 pm in CICS 208.
- Quiz 1 has been graded and the grades can be viewed in Gradescope. Historically it has been observed that Quiz 1 scores highly correlate with the performance on the course. Only 16% of students who scored 6/10 in the quiz scored more than 70% in the course. None of the students who scored less than 6/10 in the quiz scored more

than 70% in the course. Please take these statistics into consideration before signing up for the class.