Curve Fitting/Linear Regression

Skanda Bharadwaj

June 2, 2019

Curve fitting problem or the linear regression problem is solved using 2 different methods, 1.Direct error minimiztion, and 2.Bayesian approach. Results and implications from both the methods are discussed under various situations.

Contents

1 Introduction

Polynomial curve fitting or the linear regression problem is the process of constructing a curve, or a mathematical function that has the best fit to the given givent set of data points. The solution to the regression problem can be viewed in terms of minimizing the error and also interms of Basyesian approach. In the first section we solve the regression problem by directly minimizing the error function(SSE). Also, we discuss the consequences of adding the regularization term in the SSE equation. In the second section we take up the Bayesian approach and see that both the methods are almost similar.

2 Methods of Solving Linear Regression

2.1 Direct Error Minimization

Now suppose that we are given a training set comprising N observations of $x$generated from the function $\sin(2\pi x)$. The observations x are written as $\mathbf{x} = (x_1, x_2, ..., x_N)^T$ and the corresponding observations or values t is denoted as $T = (t_1, t_2, ..., t_N)^T$. We shall be fitting the data using the polynomial function of the form

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + ... + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1}$$

where M is the order of the polynomial and the polynomial coefficients $w_0, w_1, ..., w_M$ are collectively represented as $\mathbf{w}$.

2.1.1 Without Regularization

In this section, the linear regression is solved by minimizing the $error function$ that measures the error between the function $y(x, \mathbf{w})$ and the training set data points. The error function chosen here is the sum of squared errors(SSE), which is given by

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 \tag{2}$$

Minimization of 2 results in the closed form solution of w^* which is

$$W = (X^T X)^{-1} X^T T \tag{3}$$

The resulting polynomial is given by $Y = XW$

The choice of the order of the polynomial M governs the best fit to the given data. Figure 1 shows the plot for $M = 0, 1, 2, 3, 9$. It can be seen that the polynomials for $M = 0, 1, and 2$ are inflexible and the polynomial with $M = 3$ rather fits with much better generalization. It can also be observed that for $M = 9$, the polynomial fits exactly to each point which results in an overfit. Table 2 summarizes the coefficients of the polynomial $\mathbf{w}$ for orders $M = 0, 1, 2, 3 and 9$. It can be seen that the coefficients for $M = 9$ change drastically as the order increases.

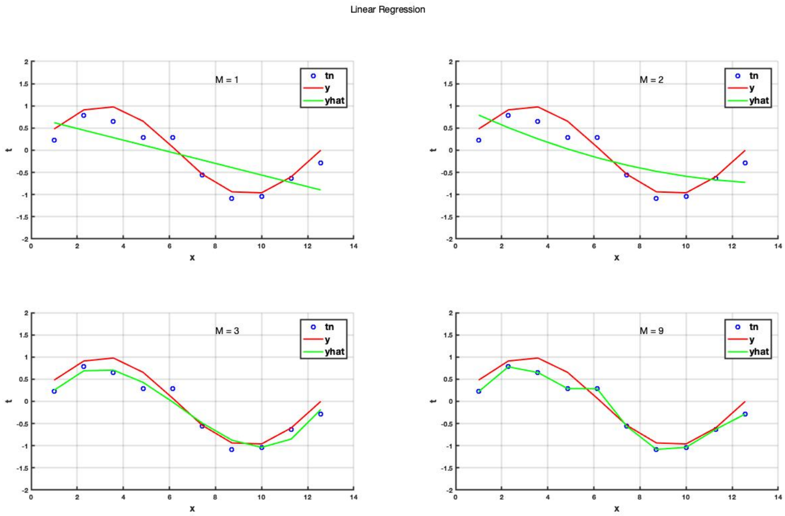
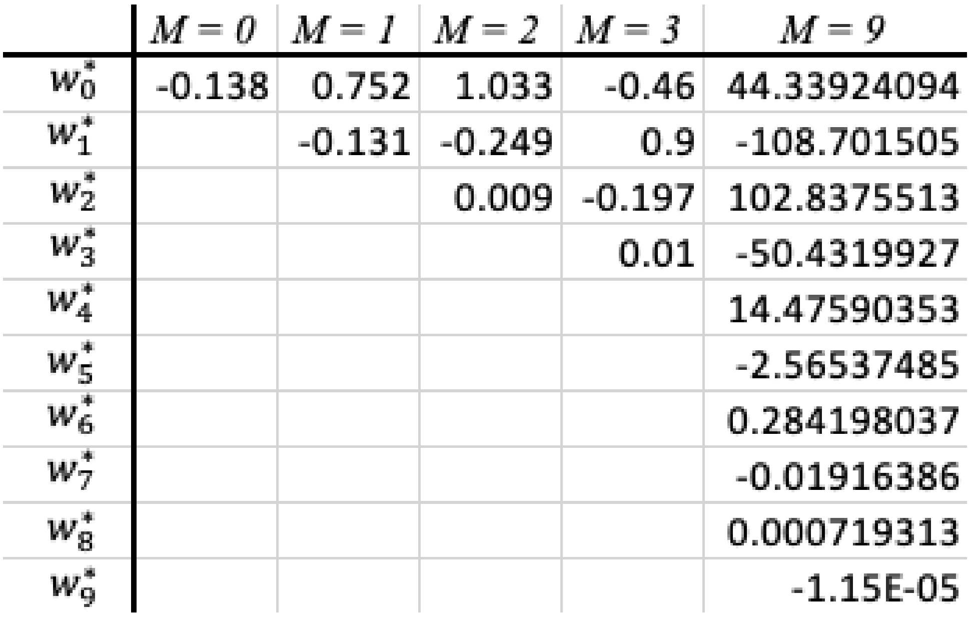


Figure 1: t_n, the scattered points represents the data (with noise), y, the red line represents the original function($\sin 2\pi x$) and $yhat$, the green line represents the estimated polynomial. Plots with $M = 1$ and $M = 2$ are inflexible and poorly generalize the data. $M = 3$ generalizes the target data better. In contrast, it can be seen that the plot with $M = 9$ fits the target data precisely at each point resulting in overfit.



	$M = 0$	$M = 1$	$M = 2$	$M = 3$	$M = 9$
w_0^*	-0.138	0.752	1.033	-0.46	44.33924094
w_1^*		-0.131	-0.249	0.9	-108.701505
w_2^*			0.009	-0.197	102.8375513
w_3^*				0.01	-50.4319927
w_4^*					14.47590353
w_5^*					-2.56537485
w_6^*					0.284198037
w_7^*					-0.01916386
w_8^*					0.000719313
w_9^*					-1.15E-05

Figure 2: W for order $M = 0, 1, 2, 3 and 9$. The values of coefficients change drastically as the order of the polynomial increases.

It can be seen that the error when $M = 9$ is zero. But yet, the polynomial very poorly generalizes the function. This is called the overfitting problem. The overfitting problem can be addressed with two different solutions. First, the number of data points can be increased. Figure 3 shows the fitting of the polynomial of order $M = 9$ for $N = 10, N = 50 and N = 100$. It can be seen that as the number of points increase the polynomial better generalizes the function. The second solution to the overfitting problem is by adding a regularization term to Equation 2, which is discussed in the next section.

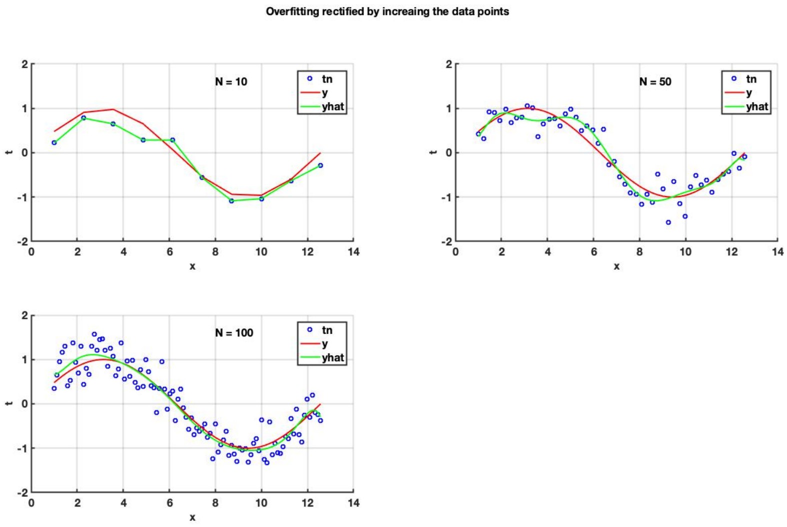


Figure 3: The polynomial of order $M = 9$ fits better as the number of data points N is increased.

2.1.2 Regularization

In order to mitigate the problem of overfitting a regularization term is added to the error function. This is given by

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}||\mathbf{W}||^2 \tag{4}$$

where $||w||^2 = w^T w = {w_0}^2 + {w_1}^2 + ... + {w_M}^2$, and the coefficient λ governs the relative importance of the regularization term compared with the sum-of-squares error term. The above equation can be written as

$$E(\mathbf{w}) = \frac{1}{2}(XW - T)^T(XW_T) + \frac{\lambda}{2}W^T W \tag{5}$$

minimizing the Equation 5, we get

$$\frac{\partial}{\partial w}E(\mathbf{w}) = \frac{1}{2}\frac{\partial}{\partial u}((XW - T)^T(XW_T) + \frac{\lambda}{2}W^T W) \tag{6}$$

$$\frac{1}{2}(-2X^TT + 2X^TXW + 2\lambda W) = 0 \tag{7}$$

$$X^T(-T + XW + \lambda W) = 0 \tag{8}$$

$$(X^TX + \lambda)W = X^TT \tag{9}$$

Therefore,

$$W = (X^TX + \lambda I)^{-1}X^TT \tag{10}$$

By substituting the value in $Y = XW$ we get the polynomial governed by λ. Figure 4 shows the difference between the two polynomials (1) without regularization and (2) with regularization where $\ln\lambda = 3$. It can be seen that the polynomial with regularization factor generalizes the function better than the one without. Figure 5 shows the variation of the error with $\ln\lambda$ for a polynomial of order $M = 9$.

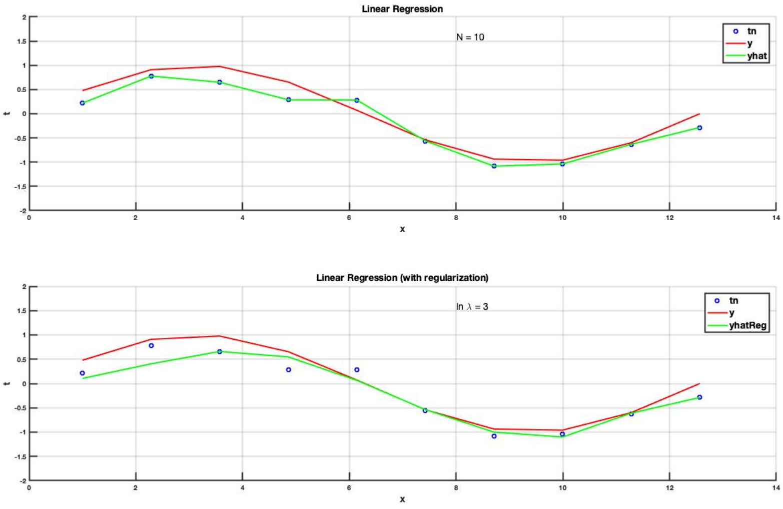


Figure 4: It can be seen that the addition of the regularization term better generalizes the function.

2.2 Bayesian Approach

Bayesian approach is another form of solving the linear regression problem from a probablistic approach. We look into the maximum likelihood approach and the maximum aposterior approach to solve the curve fitting problem.

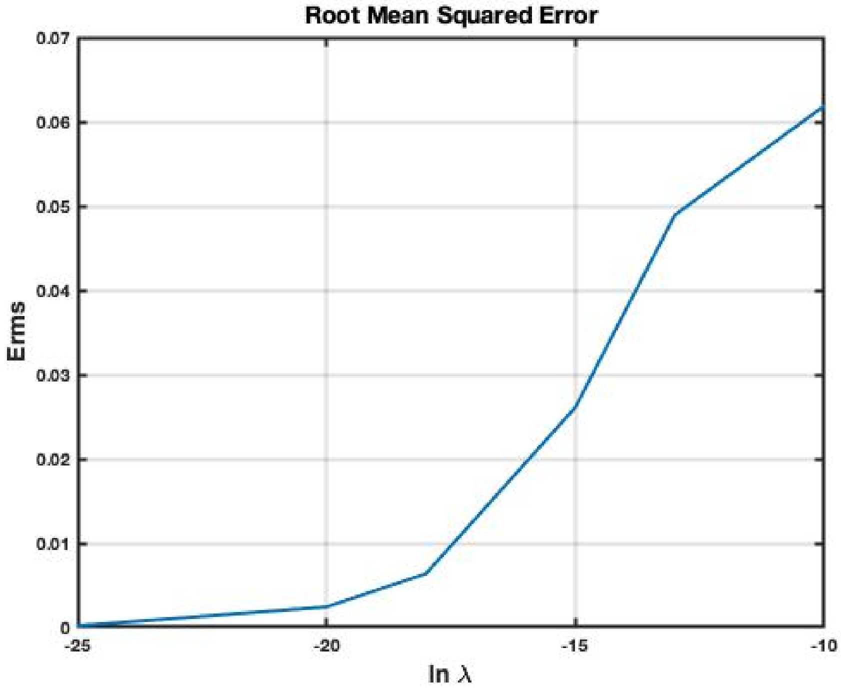


Figure 5: Graph of E_{rms} against $\ln\lambda$.

2.2.1 Maximum Likelihood Approach

In the maximum likelihood approach we consider the the bayesian probability equation given by

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \tag{11}$$

where $p(D|w)$ is called the likelihood function. In our case, the vector X can be seen ast the data given the parameters μ and σ^2 under the assumption that distribution of the data is *Gaussian*. Now, this can be written as

$$p(t|X, \mathbf{w}, \beta) = \prod_{n=1}^{N} N(t_n|y(x_n, \mathbf{w}), \beta^{-1}) \tag{12}$$

where β is the precision given by

$$\beta = \frac{1}{\sigma^2} \tag{13}$$

Taking *log* on both the sides in equation 12 we get,

$$\ln p(t|X, \mathbf{w}, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \tag{14}$$

Maximizing the above equation with respect to β

$$\frac{1}{p(t|X,\mathbf{w},\beta)}\frac{\partial p}{\partial w} = -\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{N}{2}\frac{1}{\beta} \tag{15}$$

$$\frac{1}{p(t|X,\mathbf{w},\beta)}\frac{\partial p}{\partial w} = -\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{N}{2}\frac{1}{\beta} = 0 \tag{16}$$

$$\frac{1}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 = \frac{N}{2}\frac{1}{\beta} \tag{17}$$

$$\frac{1}{\beta} = \frac{1}{N}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 \tag{18}$$

Obtaining $\mathbf{w}$ and β, we can now make predictions for new values of x. Figure 6 shows the probability distribution around the data for our curve fitting problem with $N = 10$ and $M = 3$ (The distribution is extremely insignificant for order $M = 9$.)

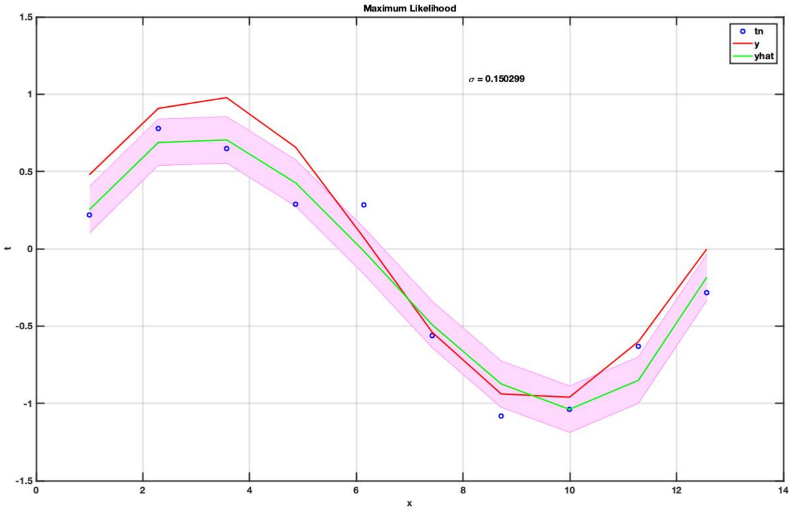


Figure 6: Probability distribution around the data using $MaximumLikelihood$ approach

It can be easily seen that maximizing likelihood is equivalent, so far as determining w is concerned, to minimizing the sum-of-squares error function. Thus the sum-of-squares error function has arisen as a consequence of maximizing likelihood under the assumption of a Gaussian noise distribution. The characteristic of maximum likelihood is that it assumes that parameters **w** is available that that the distribution of the data is found out. In constrast, maximum a posterior assumes that the data is available and computes the parameters **w**.

2.2.2 Maximum A Posteriori (MAP)

Looking at the same approach in terms of prior and posterior, the same problem can be modeled by maximizing the posterior. That is, considering prior as guassian distributed we can write

$$p(w|\alpha) = N(w|0, \alpha^{-1}I) = \frac{\alpha}{2\pi}^{(M+1)/2} exp\left\{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right\} \tag{19}$$

Using Bayes' theorem

$$p(w|X, \mathbf{w}, \alpha, \beta) \propto p(t|X, \mathbf{w}, \beta)p(w|\alpha) \tag{20}$$

Taking negative ln on both the sides we get

$$-\ln p(w|X, t, \alpha, \beta) = -\ln p(t|X, \mathbf{w}, \beta) - \ln p(w|\alpha) \tag{21}$$

Using equation 14 and 19, equation 21 can be written as

$$\begin{aligned} -\ln p(w|X, t, \alpha, \beta) = -\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi) \\ -\frac{M+1}{2}\ln\alpha - \frac{M+1}{2}\ln(2\pi) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \end{aligned} \tag{22}$$

Since the equation 22 is minimized with respect to $\mathbf{w}$, the terms without it can be neglected and we have

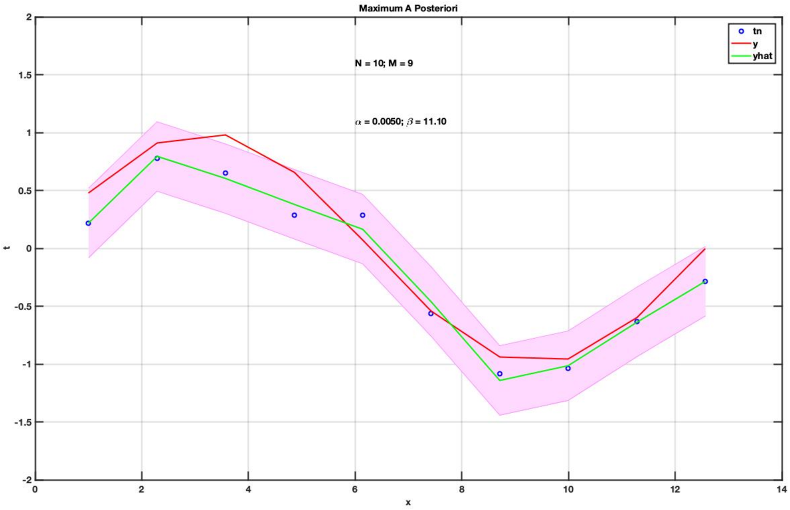


Figure 7: Probability distribution around the data using $Maximum A Posteriori (MAP)$ approach

$$\frac{1}{p(w|X,t,\alpha,\beta)}\frac{\partial p}{\partial w} = \frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w}) - t_n\}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{23}$$

which is nothing but the SSE with the regularization term $\lambda = \alpha/\beta$. Figure 7 shows the distribution found using MAP approach to our curve fitting problem for $N = 10$ and $M = 9$. It can be seen that it is very similar to Figure 4.

3 Conclusion

The project address the curve fitting problem from two approaches. First by directly minimizing the cost function and second, by taking the probabilistic or the Bayesian approach. It can be seen that both the methods are exactly similar in that the Bayesian model assumes a Guassian distribution and the SSE arises out of this assumption. We also fix the over-fitting problem either by considering more number of data points or by adding a regularization term to the SEE, which in the case of MAP we maximize the posteriori.

References

[1] C. M. Bishop (2006) *Pattern Recognition and Machine Learning*, Springer.

[2] https://en.wikipedia.org/wiki/Curve_fitting