Homework 2 Machine Learning

Marco Treglia

1 Linear Regression

Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:

One variable, denoted x, is regarded as the predictor, explanatory, or independent variable. The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

2 Mathematical step

2.1 Linear regression

Since we are interested in summarizing the trend between two quantitative variables, the natural question arises "what is the best fitting line?" In order to examine which of the two lines is a better fit, we first need to introduce some common notation:

y_i denotes the observed response for experimental unit i

x_i denotes the predictor value for experimental unit i

$\hat{y_i}$ is the predicted response (or fitted value) for experimental unit i

Then, the equation for fitting line is:

$$\hat{y_i} = w_0 + w_1 x_i$$

For computing w_0 and w_1 we need first calculate the correlation r_{xy}, the standar deviation σ and the mean μ of x_i and y_i :

$$r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \overline{x})^2 \sum_{i=1}^{n} (y_i - \overline{y})^2}}$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{n} (x_i - \mu)^2}$$

Then w_1 and w_0 are :

$$w_1 = r_{xy} \frac{\sigma_x}{\sigma_y}$$

$$w_0 = \mu_y - w_1 \mu_x$$

Another way for computing w_0 and w_1 is with the linear squares minimization , wich is computationally less expensive:

$$\arg\min_{w} \sum_{i} (y_i - wx_i)^2$$

We find w_1 and w_0 are :

$$w_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$$

$$w_0 = \frac{\sum_i y_i - w_1 x_i}{n}$$

2.2 Polynomial regression

Polynomial regression is used to fit nonlinear (e.g. curvilinear) data into a least squares linear regression model. It is a form of linear regression that allows one to predict a single y variable by decomposing the x variable into a nth order polynomial.

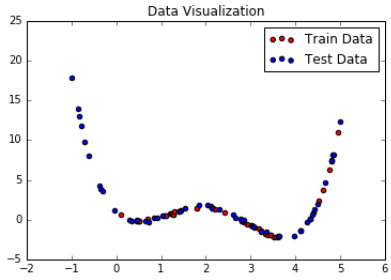
$$y = w_o + w_1 x + \ldots + w_k x^n$$

In order to evaluate how good is the predicted response we can measure the MSE (Mean Square Error). The result can be interpreted as the mesure of the accuracy of our model, the smaller it's the better way the model behaving.

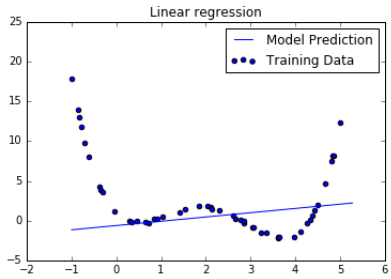
$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y_i})^2$$

3 Linear and Polynomial Regression Visualizzation

3.1 Data

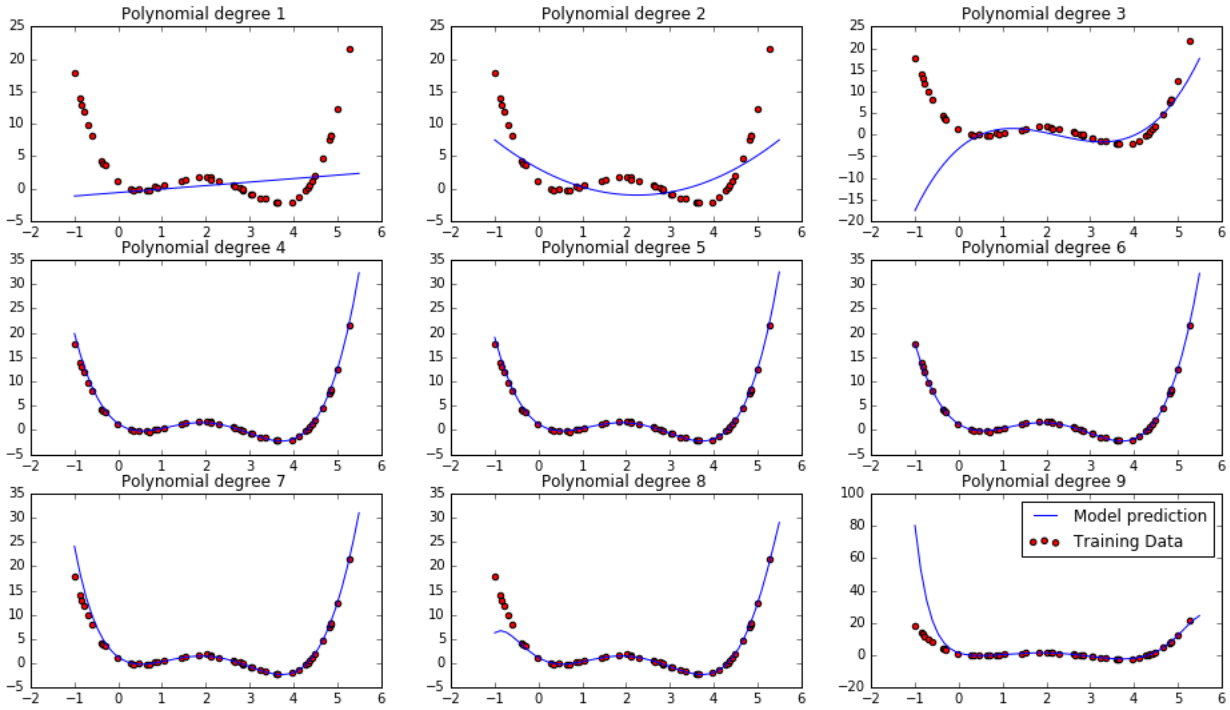


3.2 Linear Regression



MSE : 38.194

3.3 Polynomial Regression



Polynomial degree	MSE
1	38.1943
2	14.0058
3	95.6221
4	0.31226
5	0.13625
6	0.02608
7	1.98759
8	5.44563
9	157.243