Regression in Machine Learning

Apoorva Dave Follow

Dec 4, 2018 · 6 min read

Regression models are used to predict a continuous value. Predicting prices of a house given the features of house like size, price etc is one of the common examples of Regression. It is a supervised technique. A

detailed explanation on types of Machine Learning and some important concepts is given in my previous article.

Types of Regression

1. Simple Linear Regression
2. Polynomial Regression
3. Support Vector Regression
4. Decision Tree Regression
5. Random Forest Regression

Simple Linear Regression

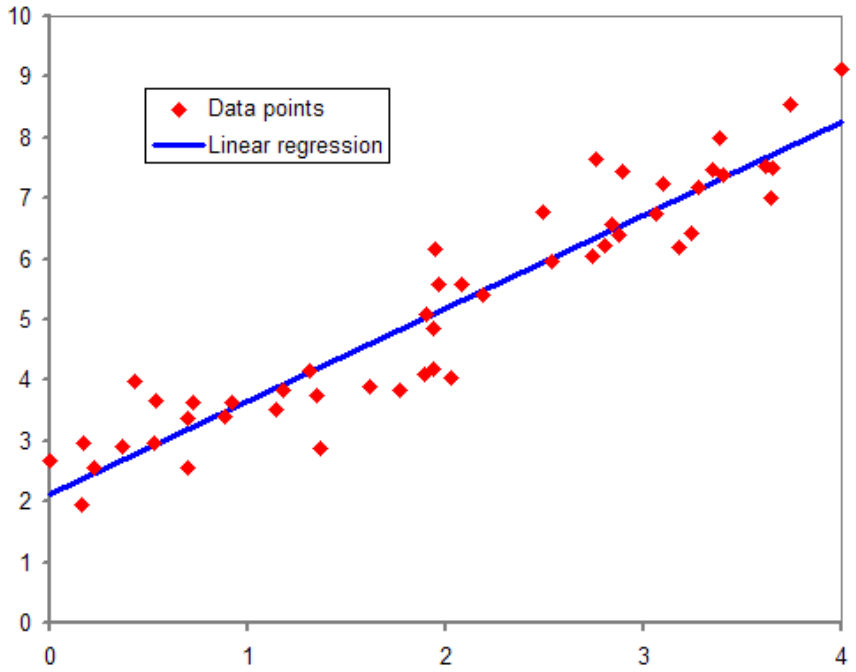
This is one of the most common and interesting type of Regression technique. Here we predict a target variable Y based on the input variable X. A linear relationship should exist between target variable and predictor and so comes the name Linear Regression.

Consider predicting the salary of an employee based on his/her age. We can easily identify that there seems to be a correlation between

employee's age and salary (more the age more is the salary). The hypothesis of linear regression is

$$Y = a + bX$$

Y represents salary, X is employee's age and a and b are the coefficients of equation. So in order to predict Y (salary) given X (age), we need to know the values of a and b (the model's coefficients).



While training and building a regression model, it is these coefficients which are learned and fitted to training data. The aim of training is to find a best fit line such that cost function is minimized. The cost function helps in measuring the error. During training process we try to minimize the error between actual and predicted values and thus minimizing cost function.

In the figure, the red points are the data points and the blue line is the predicted line for the training data. To get the predicted value, these data points are projected on to the line.

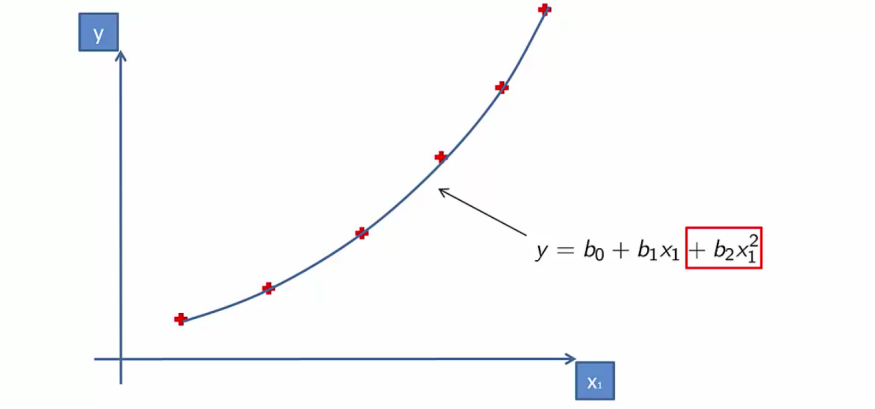
To summarize, our aim is to find such values of coefficients which will minimize the cost function. The most common cost function is **Mean Squared Error** (MSE) which is equal to average squared difference between an observation's actual and predicted values. The coefficient values can be calculated using **Gradient Descent** approach which will be discussed in detail in later articles. To give a brief understanding, in Gradient descent we start with some random values of coefficients, compute gradient of cost function on these values, update the coefficients and calculate the cost function again. This process is repeated until we find a minimum value of cost function.

Polynomial Regression

In polynomial regression, we transform the original features into polynomial features of a given degree and then apply Linear Regression on it. Consider the above linear model $Y = a+bX$ is transformed to something like

$$Y = a + bX + cX^2$$

It is still a linear model but the curve is now quadratic rather than a line. Scikit-Learn provides PolynomialFeatures class to transform the features.

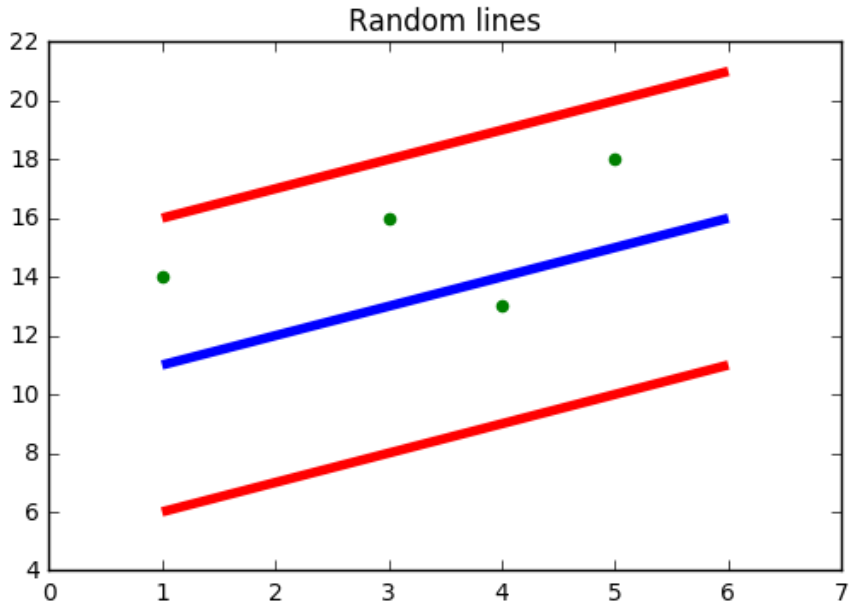


If we increase the degree to a very high value, the curve becomes overfitted as it learns the noise in the data as well.

Support Vector Regression

In SVR, we identify a hyperplane with maximum margin such that maximum number of data points are within that margin. SVRs are almost similar to SVM classification algorithm. We will discuss SVM algorithm in detail in my next article.

Instead of minimizing the error rate as in simple linear regression, we try to fit the error within a certain threshold. Our objective in SVR is to basically consider the points that are within the margin. **Our best fit line is the hyperplane that has maximum number of points.**



Data points within the boundary line

Decision Tree Regression

Decision trees can be used for classification as well as regression. In decision trees, at each level we need to identify the splitting attribute. In case of regression, the ID3 algorithm can be used to identify the splitting node by *reducing standard deviation (in classification information gain is used)*.

A decision tree is built by partitioning the data into subsets containing instances with similar values (homogenous). Standard deviation is used to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogeneous, its standard deviation is zero.

The steps for finding splitting node is briefly described as below:

1. Calculate standard deviation of target variable using below formula.

$$\textbf{Standard Deviation} = S = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

2. Split the dataset on different attributes and calculate standard deviation for each branch (standard deviation for target and predictor). This value is subtracted from the standard deviation before the split. The result is the standard deviation reduction.

$$SDR(T,X) = S(T) - S(T,X)$$

3. The attribute with the largest standard deviation reduction is chosen as the splitting node.

4. The dataset is divided based on the values of the selected attribute. This process is run recursively on the non-leaf branches, until all data is processed.

To avoid overfitting, Coefficient of Deviation (**CV**) is used which decides when to stop branching. **Finally the average of each branch is assigned to the related leaf node (in regression mean is taken where as in classification mode of leaf nodes is taken).**

Random Forest Regression

Random forest is an ensemble approach where we take into account the predictions of several decision regression trees.

1. Select K random points
2. Identify n where n is the number of decision tree regressors to be created. Repeat step 1 and 2 to create several regression trees.
3. The average of each branch is assigned to leaf node in each decision tree.
4. To predict output for a variable, the average of all the predictions of all decision trees are taken into consideration.

Random Forest prevents overfitting (which is common in decision trees) by creating random subsets of the features and building smaller trees using these subsets.

The above explanation is a brief overview of each regression type. You might have to dig into it to get a clear understanding :)

Following are a few links that might help:

Decision Tree Regression

decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g.,...

www.saedsayad.com

Support Vector Regression

Edit description

www.saedsayad.com

My next article will give an overview of different classification algorithms. Stay tuned :)

Till then happy learning!!

. . .

Cast Your Vote!

* Required

Email address *

Your email

Which Industry will see the most