greatlearning
Learning for Life

Introduction to machine learning

Linear Regression Models -

a. Before we generate a model, we need to understand the degree of relationship between the attributes Y and X

b. Mathematically correlation between two variables indicates how closely their relationship follows a straight line. By default we use Pearson's correlation which ranges between -1 and +1.

c. Correlation of extreme possible values of -1 and +1 indicate a perfectly linear relationship between X and Y whereas a correlation of 0 indicates absence of linear relationship

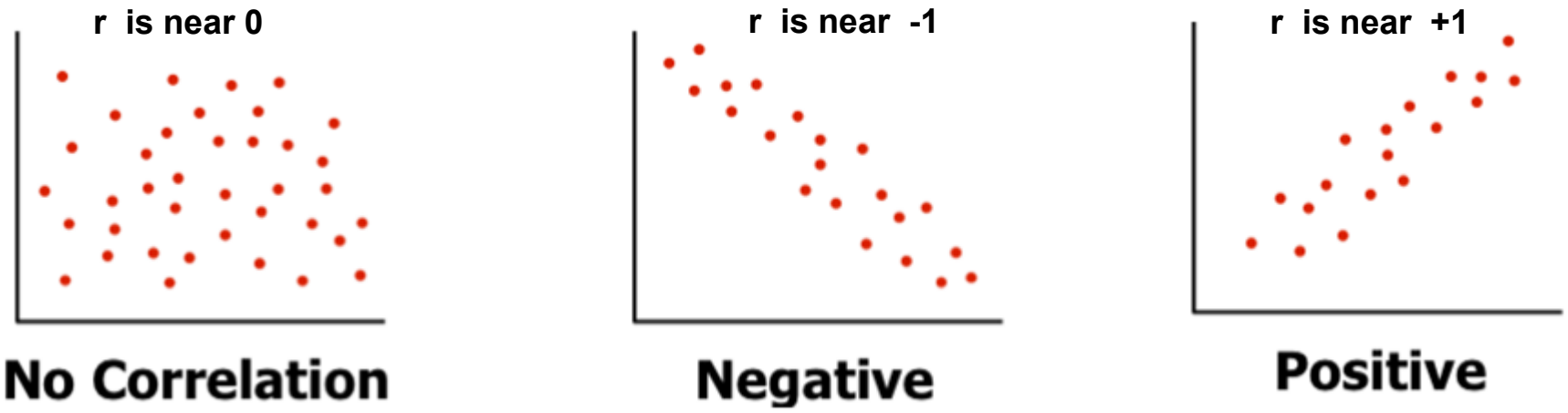
   I. When r value is small, one needs to test whether it is statistically significant or not to believe that there is correlation or not

Introduction to machine learning

Linear Regression Models -

d. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



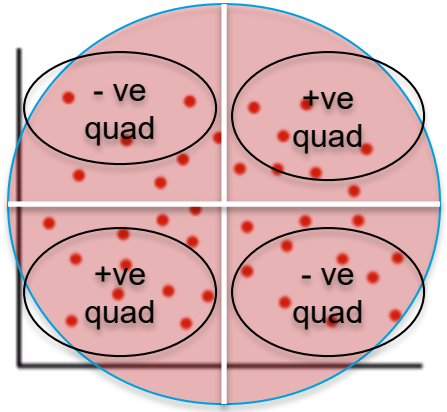
e. Generating linear model for cases where r is near 0, makes no sense. The model will not be reliable. For a given value of X, there can be many values of Y! Nonlinear models may be better in such cases

Introduction to machine learning

Linear Regression Models (Recap) -

f. Coefficient of relation - Pearson's coefficient p(x,y) = Cov(x,y) / (stnd Dev (x) X stnd Dev (y))

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$



$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = 0$$

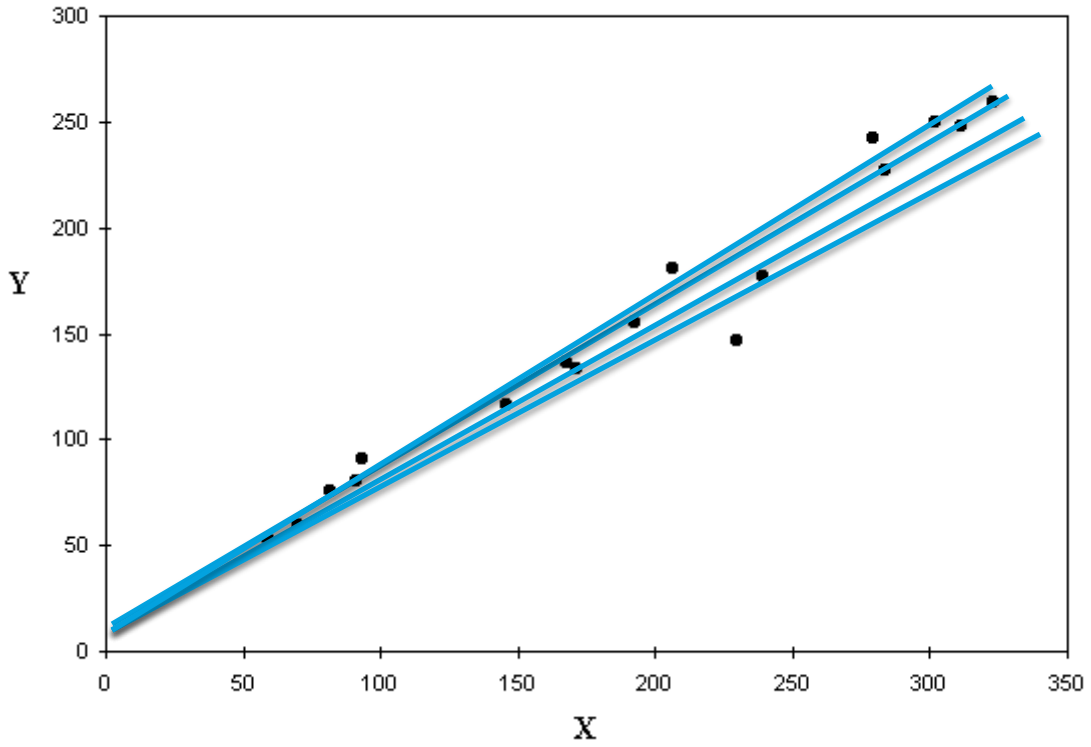
http://www.socscistatistics.com/tests/pearson/Default2.aspx

Introduction to machine learning

Linear Regression Models -

g. Given Y = f(x) and the scatter plot shows apparent correlation between X and Y
Let's fit a line into the scatter which shall be our model

h. But there are infinite number of lines that can be fit in the scatter. Which one should we consider as the model?



i. This and many other algorithms use gradient descent or variants of gradient descent method for finding the best model

j. Gradient descent methods use partial derivatives on the parameters (slope and intercept) to minimize sum of squared errors

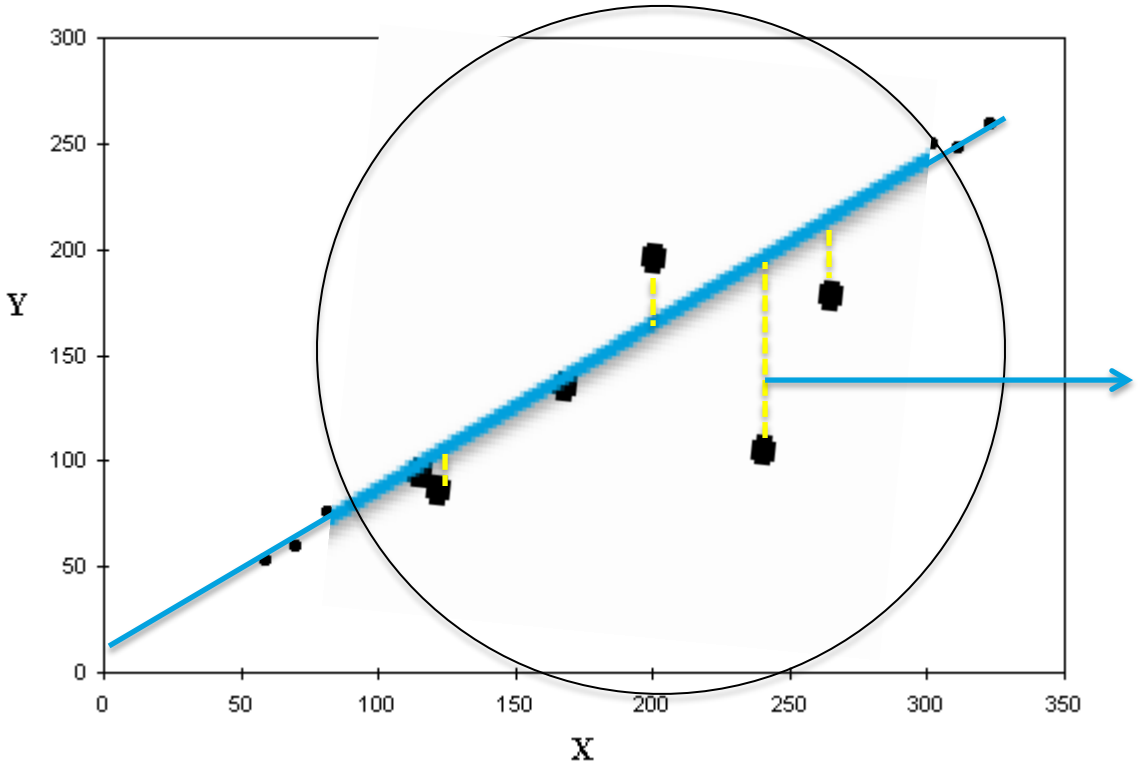
Introduction to machine learning

Linear Regression Models (Recap) -

k. Whichever line we consider as the model, it will not pass through all the points.

l. The distance between a point and the line (drop a line vertically (shown in yellow)) is the error in prediction

m. That line which gives least sum of squared errors is considered as the best line



Error = (T – (mx + C)

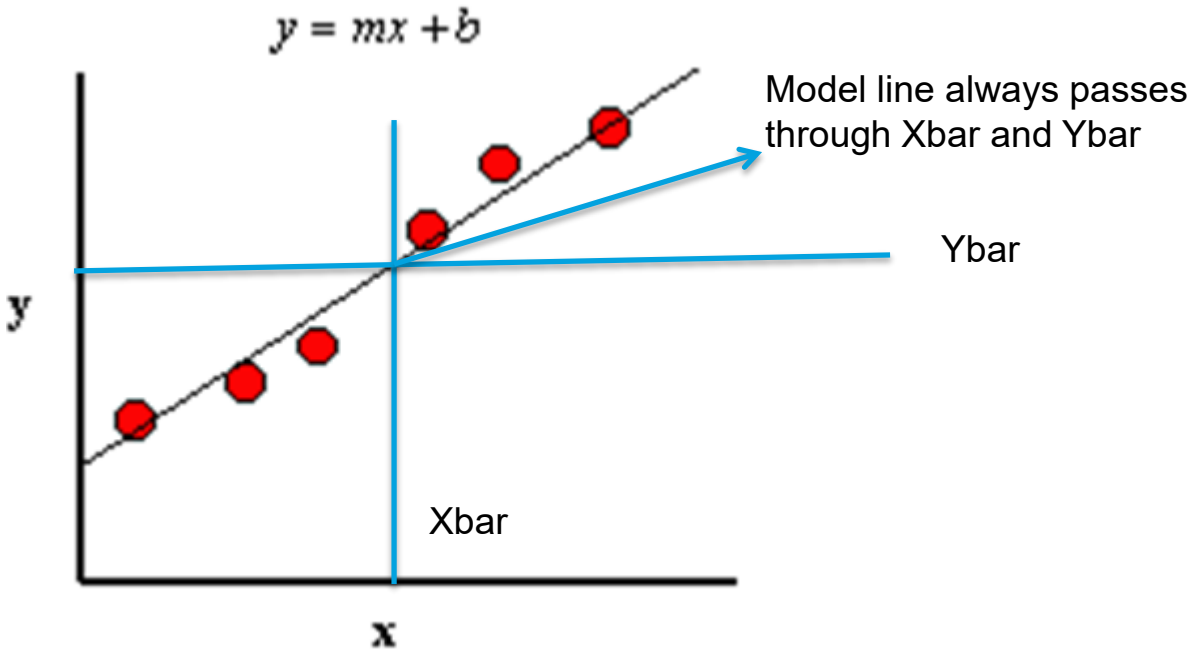
Sum of all errors can cancel out and give 0

We square all the errors and sum it up. That line which gives us least sum of squared errors is the best fit

Introduction to machine learning

Linear Regression Models -

n. Coefficient of determinant – determines the fitness of a linear model. The closer the points get to the line, the R^2 (coeff of determinant) tends to 1, the better the model is



Introduction to machine learning

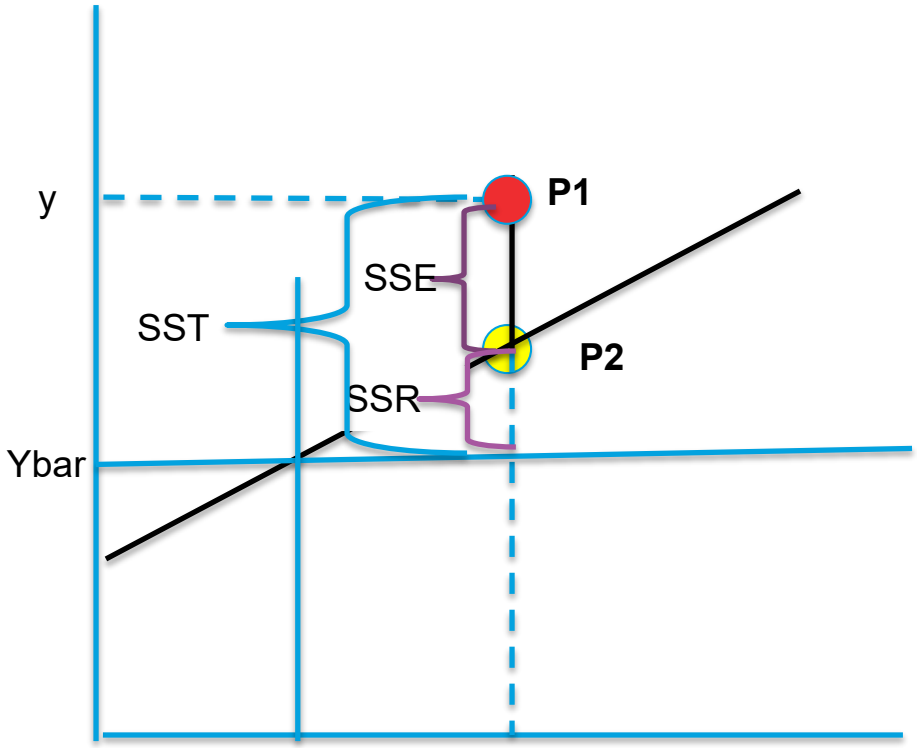
Linear Regression Models -

o. Coefficient of determinant (Contd…)

I. There are a variety of errors for all those points that don't fall exactly on the line.

II. It is important to understand these errors to judge the goodness of fit of the model i.e. How representative the model is likely to be in general

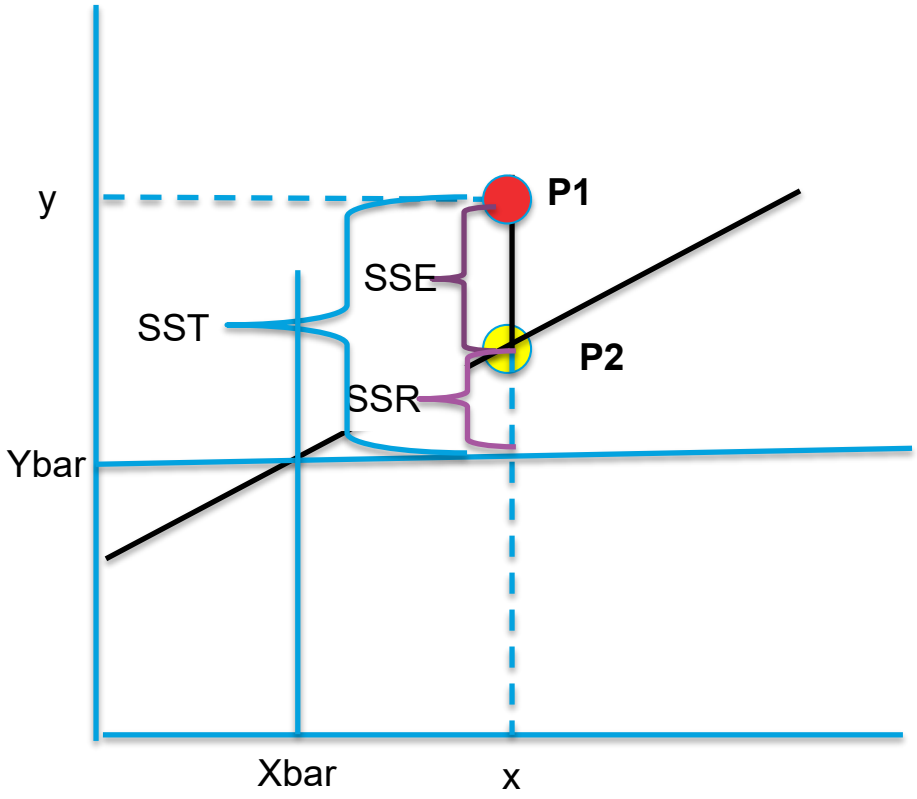
III. Let us look at point P1 which is one of the given data points and associated errors due to the model



1. P1 – Original y data point for given x
2. P2 - Estimated y value for given x
3. Ybar – Average of all Y values in data set
4. SST – Sum of Square error Total (SST) Variance of P1 from Ybar (Y – Ybar)^2
5. SSR - Regression error (p2 – ybar)^2 (portion SST captured by regression model)
6. SSE - Residual error (p1 – p2)^2

Introduction to machine learning

Linear Regression Models -



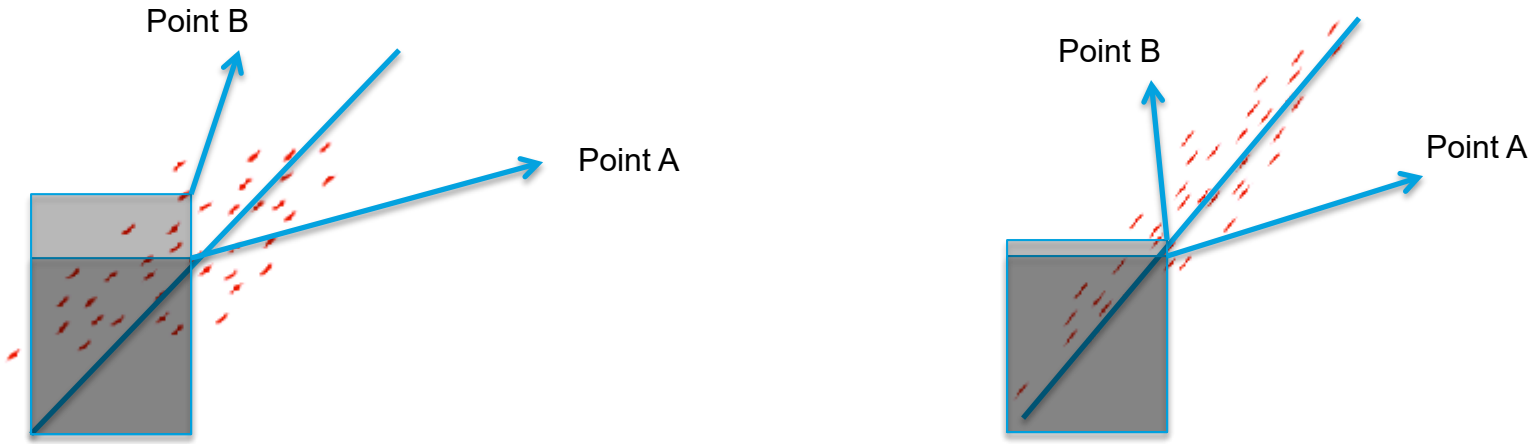
p. Coefficient of determinant (Contd…)

1. That model is the most fit where every data point lies on the line. i.e. SSE = 0 for all data points
2. Hence SSR should be equal to SST i.e. SSR/SST should be 1.
3. Poor fit will mean large SSE. SSR/SST will be close to 0
4. SSR / SST is called as r^2 (r square) or coefficient of determination
5. r^2 is always between 0 and 1 and is a measure of utility of the regression model

Introduction to machine learning

Linear Regression Models -

q. Coefficient of determinant (Contd…) -



In case of point “A”, the line explains the variance of the point

Whereas point “B” the is a small area (light grey) which the line does not represent.

%age of total variance that is represented by the line is coeff of determinant