03 Regression

Thursday, 4 March 2021   12:15 PM

Stats v ML

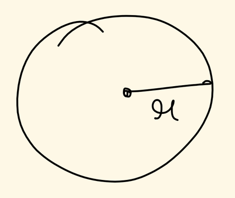
- ML primarily concerns with minimizing the error of a model or to make the most accurate prediction possible at the expense of explainibility.
- we use ideas & algorithm from different fields like computer science, stats etc. & we use them towards their ends.

Now, Linear Regression was developed in the field of stats and has been extensively used to study relationship between i/p & o/p numerical variables. In Machine learning it has a similar use as well.

Linear Regression is a linear model
it assumes relationship b/w i/p variable (x) & o/p var (y).
i.e y is calculated from a linear combination of x.

(1) Relation Based Prediction

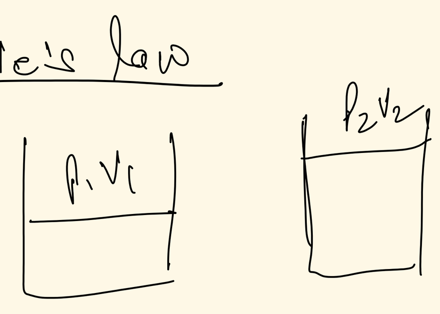
Eg:-



$\boxed{C = 2\pi r}$
$C \propto r$

or

Boyle's law



$= P_1V_1 = P_2V_2$

} Rigid Physics

Provide example of linear & Non-linear Models.

Flexible

(1) Height of offspring vs the height of parents.
Can draw a relationship b/w the height of father and son.

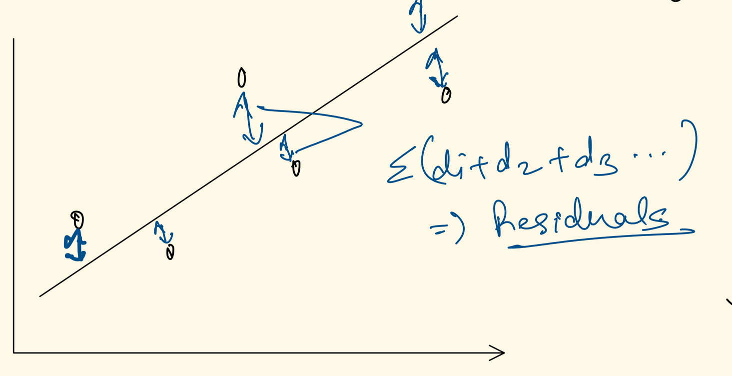
while many would follow a general trend, there would certainly be cases where it won't. } Source of error

Regression :-

Regression means that in general we can express the height of son and fathers using an equation involving the height of father and son.

$R(h_f, h_s) :\Rightarrow h_s = ah_f + b$
or $h_s = b + ah_f$ } linear Regression

Linear Regression :- The idea is to fit a line that's nearest to all values at once. The values of this line is nearest to the actual value and we take this as the forecast or estimate.



} Linear Regression aims to find the smallest sum of Residuals that's possible for the dataset

R^2 :-> goodness of fit {0, 100%}

-> It indicates the variance in the dependent variable that the independent variable explains.

[0-1]

$R^2 = \frac{\text{Variance explained by the model}}{\text{Total Variance}}$

Caution :- ① Cannot be used to evaluate the coefficient estimates or bias in prediction. Hence look at the plot

② A good model can have low R^2 value and a bias model can have a high R^2.

Ordinary Least Square / Multivariable regression

Simp. Linear. Reg :-> $\boxed{y = ax + b}$ where {a, b} -> Coefficients

{ I -> input
I -> output }

y: dependent
x: Independent

Ordinary Least Sq : more than 1 input. 1 output.

$$\boxed{y = a_1x_1 + a_2x_2 + a_3x_3 + b}$$

Seeks to minimize the sum of the square residuals.

i.e ① we fit a line

② we take the distance of each point from the line

③ we square it and sum it all. } Minimize it

$\underline{OLS} = \Sigma(d(\text{point} \& \text{line}))^2$

The approach treats the data as matrix and uses linear algebra to find optimal values for the coefficients

-> Disadvantage :-> {Memory hogger}

-> Advantage :-> Very very fast.