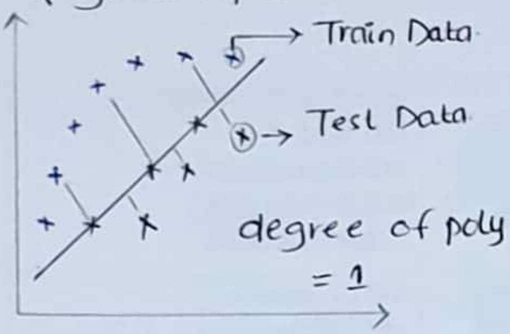
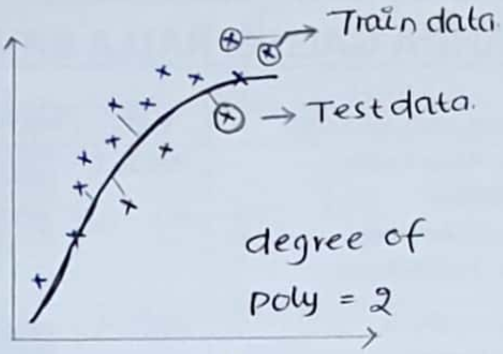
Overfitting & Underfitting :-

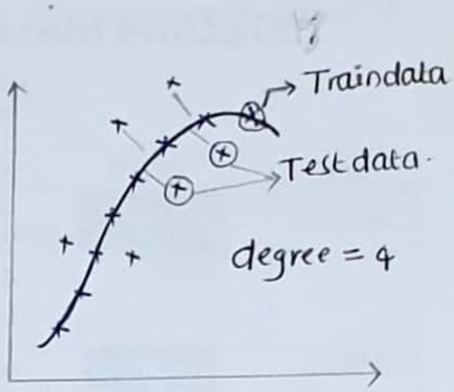
Regression :-



(a)



(b)



(c)

Under fitting → High Bias, High Varience

Best Model → low. B., low. V.

Over fitting → low Bias, High Varience

Case - I :- from the above fig we draw a best fit line or Curve Using linear regression where we use polynomial degree, in fig (a) we apply degree = 1 its act like a Simple linear regr. and draw a line on training data points when we use Scatter plot graph., after it we Caliculate R^2-Error, if we observe the error will be high becouse the distance between the actual value and predicted value is more ie error rate will be more. the model gives us high error.

Note :- In case of training data it self the model gives us high error. ie "high bias" is called as "Under fitting" issue.

Case - II :- if we observe graph (c), where we applying degree = 4 its form a Curve which is best fitted to all the training datapoints. So from this we can Say On training data there is no error because Curve fitted good., if we have a new test data points Come in the graph then that makes Some distance from Curve ie errors formed in model. here Our aim is to fit the Curve best, from that recorded low error from this we can Say Accuracy of trn data is more, but for test data.

is low becouse error form in test data. is called "Over fitting" ②

In case - i the Accuracy of training data is low (huge Errors) as well

the Accuracy of a test data also low (Errors).

Here Our aim is to be get high Accuracy in case of train data

as well test data (ie prevent the errors)

if we Observe the figure (b) i think its a best fitted Curve at

polynomial degree = 2 and where has no errors ie minute errors in train

data aswell test data. So i prefer this model, ~~to~~ becouse this gives

high Accuracy in train data, test data. and error rate should be low

as compare to Other two models where overfitting & Under fitting problem

effected to the model.

if we observe model 2 that gives us low bias and low Varience

if in Underfitting case that gives high bias and high varience ie that

the error rate will be high in train data called as high bias Similarly, if

error rate will high in test data. Called as high Varience.

In case of Overfitting. that gives low bias and high varience ie the error

rate in train data is low but error rate in test data is high.

Varience :- if your model changes a lot with a Small change in

the training data this means the model has "High Varience"

the Issue with high varience is that the model becomes very Complex.

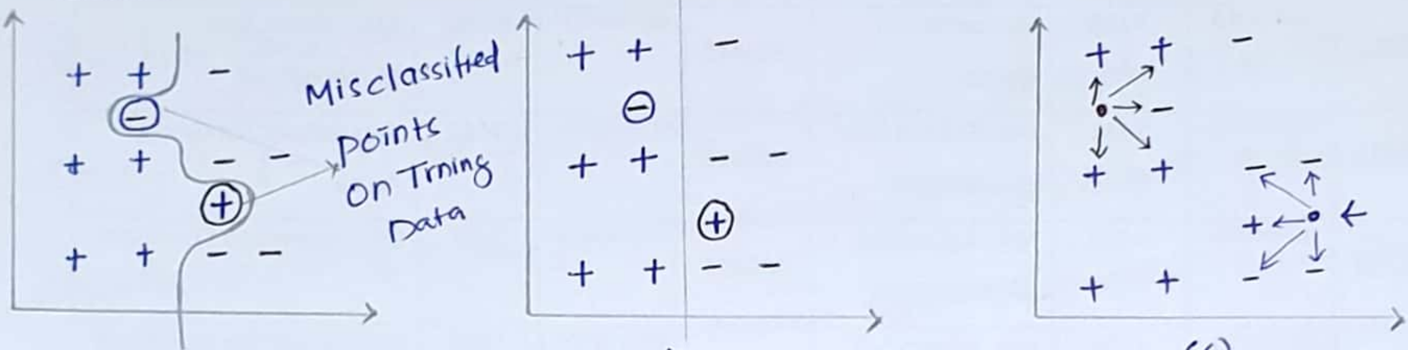
to avoid Complexity we have to choose Simple model.

Occam's Razor theorem Says that try to learn Simple models but not too Simpullar models.

Reason behind choose Simple model is that going to be more "generalize" On test data (ie generalizable).

Note :- In Regression we classify the model as overfit or Under-fit based on "Errors" in training data aswell in test data, Similarly In classification we classify based On misclassified data points in training data aswell test data.

Classification :-   logistic Reg.   K.N.N. Classification



(a)   (b)   (c)

Overfitting — low bias, highvarience

Best model — low bias, low var.

Underfitting — H. bias, H. var.

from the above figure (a) if we plot a line which is best Classify or Separate a data points is Consider as classification task if we plot a line across this data. we get this two misclassified points on graph for Suppose if we have a new data point (or) test data Comes to In this this will get more change in data set ie when train data. we have less classifier in Correct point, when ever we consider test data points the change in Classifier is more then we can Say that the model is "Overfitting"

This overfit model Containing low bias and high Varience

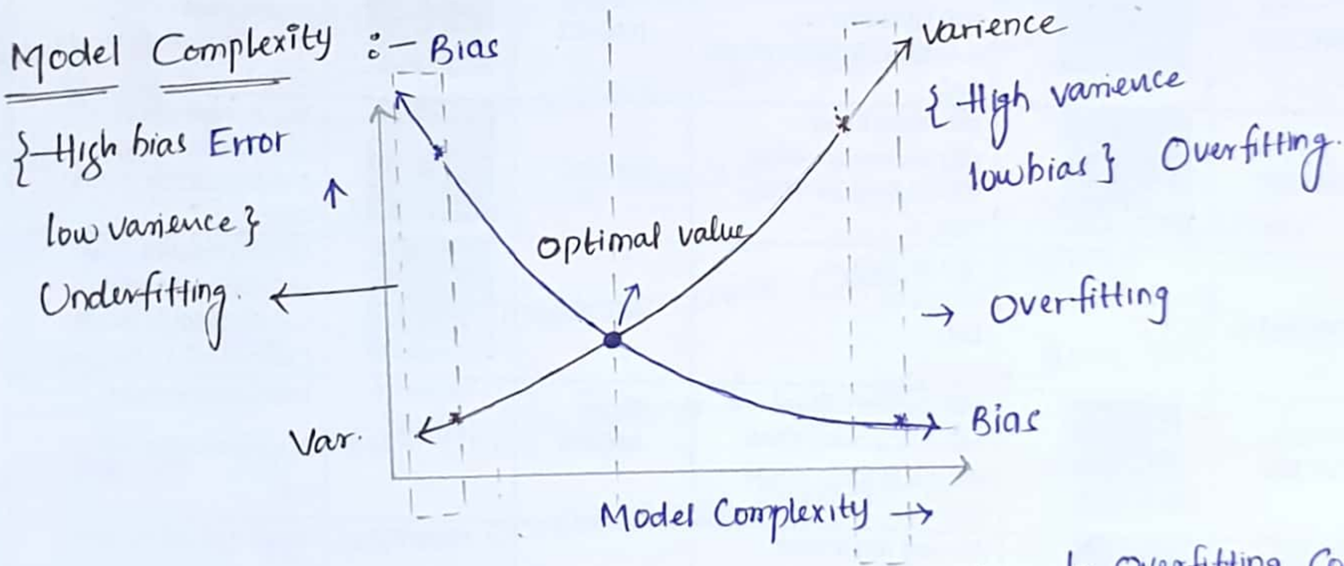
Case - II :- if we observe a graph (b), that line Separates the data points even in that we get Some mis classified points where we applying logistic Regression classifier. if we get low mis classified points on Separate data in training data. as well when we apply ~~a~~ test data to this model then also if we get Some minute mis classified points in test data. that are changeble. then there is no more impact of test data. (ie not classified more points) then we can Say that is the "Best model", this model Containing low bias and low Varience.

Case - III :- if we observe figure (c), there we apply KNN classification algorithm to classify a data point ie belonging to which Categiry of that point In K.N.N the term say that K- nearest points Observation where the nearest points are which group or belonging to which Catogyry are Obtained more that point should be that categiry ie the test point have also Similar properties. here depends on 'k' value the categiry of point ~~a~~ classified

if we observe the graph (c) if we take any new test data point on graph from the value if we Consider "k = 4" then Compare all data points for Suppose we get (3+ve) (1-ve) point then we can Say the new point is "+ve" point but if we know that the point is -ve (ie the test point) but in k nearest points all are positive then we can say that the point is mis classified. here 'k' value is more Important.

Let Consider we have total training points are '13' and we mentioned 'k = 13' that means all points Comes under Observation. then we find the nearest points distance of all 13 data points. after that we Sort ~~it~~ all data points. but In this we Consider k value as total training data points So no need of Sorting if we classify the all data points here +ve's are 7 and -ve's are 6 So the new data point should be Consider as a +ve point this is big problem because the model is dumb model (ie because of highest value as "+ve" class, So every time when ever new data point Comes we classify that as +ve point.) where high misclassified data points Comes in training data as well test data. So then we can easily Say that this model should be overfitting model

Overfit model Consist of ~~High~~ varience as well High bias.



So from this graph we can say that On training data In Overfitting Contain high varience, Similarly Underfitting Contain ~~high varience~~ low varience for test data In Overfitting Contain low bias, Similarly On Underfitting Contain high bias as shown graph.

Short Notes On Overfit vs Underfit :-

Overfit vs Underfit

Overfit
↓
On train Data we have high varience
↓
this are very Complex models.
↓
train error is low
Test error is high

Underfit
↓
High bias
↓
Very Simple model
↓
train error is high
test error is high

Bias Varience Tradeoff :- it is also called as "Optimal value point" as shown in figure "model Complexity", actually what this point do is when ever this problems come Underfitting and Overfitting. we unnecessarly have high varience and low bias in Complex models, where we have high bias and low varience in Simple models., this both high varience and high bias going to kill model (ie large error rate) and it never be generalizable on test data. In order to reduce (ie maintain low varience, low bias) we Use this Optimal value point. (or) trade off.

Regularization :- In order to treat Overfit & Underfit problems we ~~use~~ do Regularization, is nothing but a technique used for tuning the function by adding on additional penalty term in the error function. In other words, this is a form of regression, that regularize or shrinks the coefficent estimate towards zero and this discurages learning a more Complex of Simple model, So to avoid the risk of overfitting.

<u>Note</u> :- All Machine learning Algorithms will have some regularization Steps inbuilt. (ie we have to take care of this terms)

Let Consider linear Regression & logistic regression here we regularize Some terms in this algorithem that prevent the Issue of Overfitting and Underfitting.

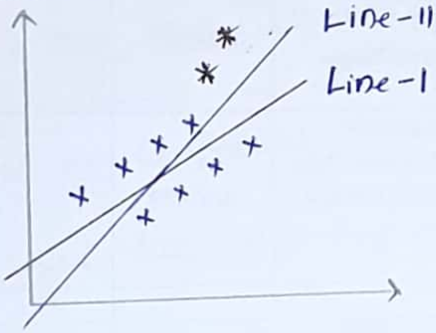
Consider the linear reg eqn — $\boxed{m^*, c^* = \arg_{m,c} \text{Min} \sum (y_i - (mx+c))^2}$

from the eqn we replace m with ω, — we get eqn — $w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_0 \end{bmatrix}$

$$\boxed{\omega^* = \arg_w \text{Min} \sum (\underbrace{y_i - (\omega^T . x_i + w_0))^2}_{\text{Cost function}}}$$

Here we minimize the Errors from the eqn

So in regression process we fit a line that best fit the data points.



Overfitting Model

from the above graph assume intially there is no lines, we plot data points on graph and making a line-I that best fit the all data. points, for Suppose if we Consider two more points (test points) which away from all the points (ie Outliers) In this Situation line-I is not a best fit line So we have to again plot a line that best fit the data points. let Consider line-II that best fit all the data points. (ie minimize the error) from this line making model will be change if we observe this

We can say for a small change in train data the entair

model will be changed (ie high varience) effected. Simply we

can say our linear regression model has overfitting problem.

In order to prevent this problem we use Regularization. this

mainly effect on errors. from the linear eqn we have to cost function.

So our aim is to minimize cost function. So simply we can represent –

$$\boxed{w^* = arg_w \ Min \{ Cost function \}}$$

In order to reduce cost function In linear regression we have to use

'Gradient Descent' parameter.

So now we have to represent regularized term In order to min error

$$\boxed{w^* = arg_w \ Min \{ Cost function + Reg.(w) \}}$$

where $w \rightarrow$ coefficent (m,c) in linear Regression.

So now here we used two diff regularized terms for linear Regression

they are 1) L1 Regularized also called as lasso Regularizer

2) L2 Regulariser also know as Ridge Regularizer.

this both terms comes from (L_1, L_2) Manhatten, Euclidian distance

In L1 Regularizer we do – $\Sigma |w_i|$ absolute.

In L2 Regularizer we do – $\Sigma (w_i)^2$ because M.S.E

when ever we use this regularizers we have to multiply this

term with lambda denoted as (λ), it is a Inbuilt function in

regularizers. like $(L_1 \$ L_2)$

Now finally Eq's will be —

linear reg eqn with L_1 Rglzr — $\boxed{w^* = \arg_w \text{Min}\{\Sigma(y_i - (w^T.x_i))^2 + \lambda \underbrace{\sum_{i=1}^{n} |w_i|}\}}$

lasso Rglzr

linear reg with L_2 — $\boxed{w^* = \arg_w \text{Min}\{\Sigma(y_i - (w^T.x_i))^2 + \lambda \sum_{i=1}^{d} (w_i)^2\}}$

Ridge Regularizer

Logistic Regression :-

Now we consider logistic eqn — $\boxed{m^*, c^* = \arg_{m,c} \text{Max}\{\Sigma(y_i * (mx_i + c))\}}$

In this eqn we have a problem with Outlier, because of this we get incorrect result; to avoid this use Sigmoid function, when apply this function the eqn will be — $\boxed{m^*, c^* = \arg_{m,c} \text{Max}\{\Sigma \frac{1}{1+e^{-x}}\}}$

Now we consider the $(y_i * \hat{y})$ as a function — 'f' then —

$$m^*, c^* = \arg_{m,c} \text{Max}\{\Sigma \frac{1}{1+e^{-f}}\} \Rightarrow \text{Can be} \Rightarrow \exp\{-f\}$$

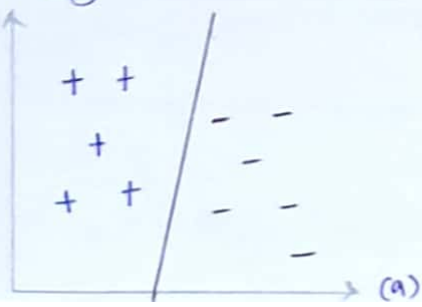
Now from eqn put 'f' value $\Rightarrow m^*, c^* = \arg_{m,c} \text{Max}\{\Sigma \frac{1}{1+\exp\{-y_i * (mx_i + c)\}}\}$

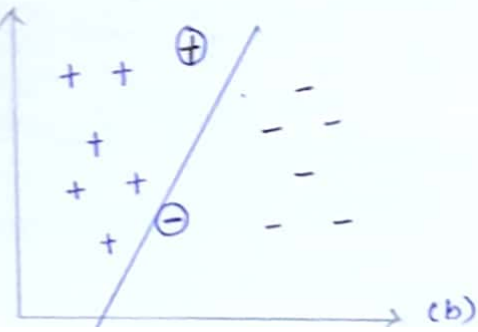
So when we minimize the value of 'f' then only we get Max result

In Order to do it we do inverse of it — $\max \frac{1}{f} = \min f$

finally We get — $\boxed{m^*, c^* = \arg_{m,c} \text{Min} \sum_{i=1}^{n} \{\underbrace{1 + \exp\{-y_i * (mx_i + c)\}}_{\text{cost func}}\}}$

This eqn represent logistic Regression after treatment of Outliers with use of Sigmoid function, which tries to classify the positives to Negatives





from figure (a) we know that the line which is best Separate the data from positives to negatives. but what happend when new data point (test data) Comes, then we have to drastically line formation changes according to new data points if we Observe fig (b) the line angle will be change as compare with fig (a). and in fig (2) we Correctly classify the data points Intially which are missclassified data points.

" Our aim is to reduce the missclassified points In logistic Regression "

So that the best Separated line fig (2) reduce the mis classified data points. from this Separation line making model will be changed. if we Observe this we Can Say for a Small change in train data the entain model will be changed (ie high varience) effected So we Can Say this model has Overfitting issue., to Solve it we use regularizer methods.

In logistic Regression we have a Cost function, So Our aim is to Correctly classify the data points, the eqn will be represented as —

$$\boxed{w^* = \arg_w \text{Min}\{\text{Cost function}\}}$$

So Now we represent regularized ~~term~~ term Inorder to Current Classify.

$$w^* = \arg_w \text{Min}\{\text{Cost function} + \text{Reg}(w)\}$$

In logistic also we used L_1 & L_2 Regularizers. —

Logistic reg. with L_1 Regulr. — $w^* = \arg_w \text{Min}\{\Sigma(1+\exp\{-y_i * (mx_i+c)\}\} + \underline{\lambda \Sigma_{i=1}^{n} |w|}\}$ → lasso Reglzr

L_2 Reglzr — $w^* = \arg_w \text{Min}\{\Sigma(1+\exp\{-y_i * (mx_i+c)\}) + \underline{\lambda \sum_{i=1}^{d} (w_i)^2}\}$

Ridge