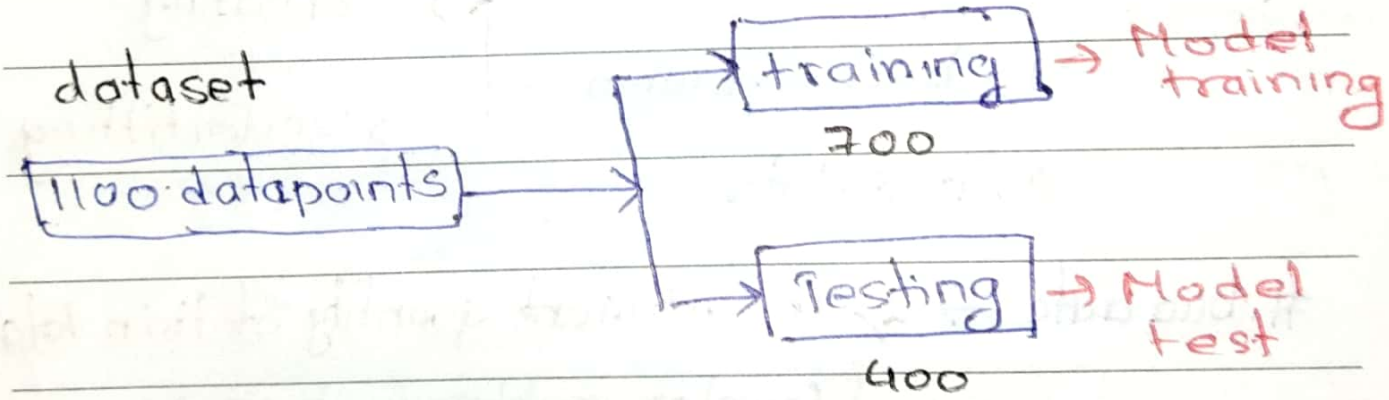
the model.

Overfitting and Underfitting:



Overfitting

Train → Very good accuracy ⇒ low bias
(90%)

Test → Bad accuracy ⇒ High variance
(50%)

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Underfitting:

TRAIN → Model accuracy is low → High bias

TEST → Model accuracy is low or high → Low or high variance

We can solve this issue by performing hyperparameter tunning or by increasing no of dataset.

Main challenges in ML:

1) bad algorithm → overfitting, underfitting

2) bad data

bad data → 1) Insufficient quantity of train data.

(complex problem such as image or speech recognition may need millions of data unless we use existing model)

2) Non representative training data.

(missing some significant values)

3) Poor quality data:

→ training data full of error, outliers noise. So hard to detect underlying patterns)

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

4) Irrelevant features:-

(garbage in, garbage out)

A critical part of success of the ML project is to come up with good set of features to train on. This process is not as Feature Engineering. Involves

1) feature selection → selecting most import features among features

2) feature extraction → Combining existing features to produce a more useful one like dimensionality reduction algorithm.

3) Creating new features by gathering new data.

Instance Based Vs Model based

Instance based → calculate vector distance with neighbours, select any N number of nearest neighbour, based on the probability of similarity, we make decision.

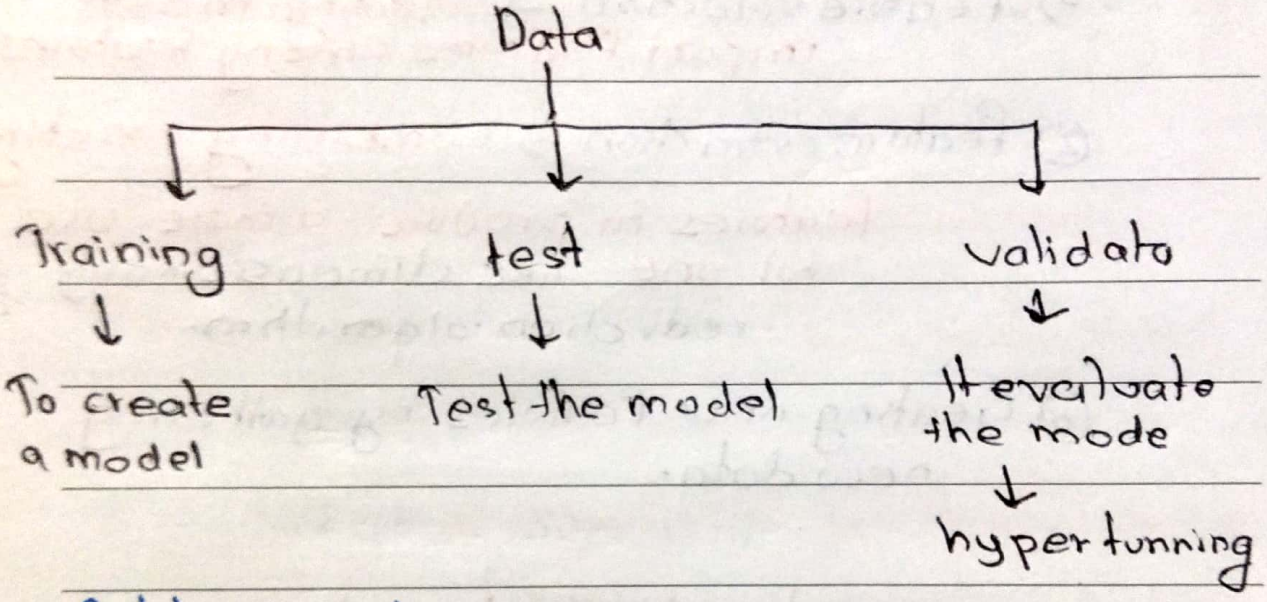
Memo No. _____

Mo Tu We Th Fr Sa Su

Date / /

Model based → A model which is decision function is generated and based on that function, prediction is made.

Train, Validation and Test sets



Batch Vs Online learning

Batch :- System is uncapable of learning incrementally. It must be trained using all the available data.

Online:- System can be trained feeding data sequentially.

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Regularization

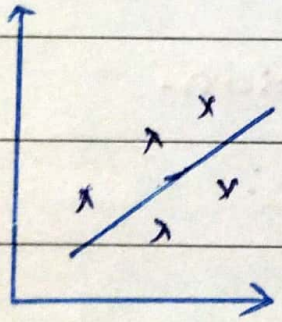
→ to avoid overfitting

Regularization technique:-

1) Ridge/L2 Regression:-

→ One of the method of regularization technique which the data suffers from multicollinearity. → high correlation between two or more independent variable in multiple regression model.

Overfitting ⇒ low bias and high/low variance.



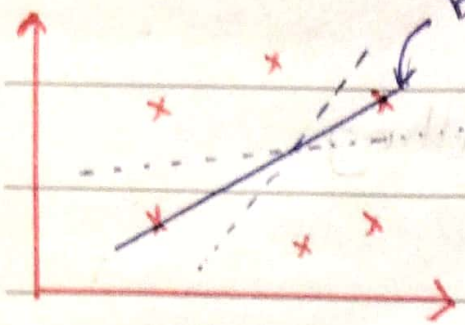
- If the data (test data) is near to best fit line, then performance will be good. (Low bias)
- If the test data is far away from best fit line, then performance will be bad. (high variance)

AIM: To reduce overfitting

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /



→ we create multiple lines to improve performance of test data.

Cost function

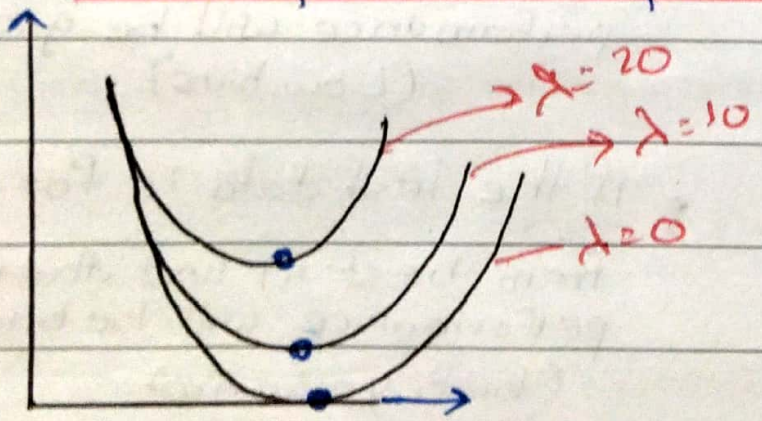
$$\text{Cost function} = \frac{1}{m}\sum_{i=1}^{m}\left\{h_\theta(x)^i - y^i\right\}^2 + \lambda\sum_{i=1}^{n}(\text{slope})^2$$

here, λ = hyperparameter.

slope: slope of different lines.

if $\lambda = 0$; ⇒ cost function is same as line or regression's cost function.

Relationship between slope and λ:



"shifting towards zero" → global minima

Mo Tu We Th Fr Sa Su

Memo No. _____

Date / /

Global minima gets shifted towards left with increase in λ.

Cost function = $0 + (\text{slope})^2 \cdot \lambda$

: +ve ↓↓

Change θ value to create another best fit line

$$\lambda \propto \frac{1}{\text{Slope}}$$

#Inversely proportional relationship.

λ = It make sure that our line doesn't overfit.

$\lambda \geq 0$ is a complexity parameter that controls the amount of shrinkage.

the larger the value of λ, the greater the amount of shrinkage.

The coefficient are shrink towards zero.

* θ value never becomes zero!

Memo No. _____

Date / /

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3$

$= \theta_0 + 0.95 x_1 + 0.82 x_2 + 0.10 x_3$

∴ Ridge regression is used to introduce bias to the data in order to generalize the data and increase bias.

This is useful if you don't have much training data.