→ Where @ K=n, It is Called Underfitting because, The Classifier is Underworking because its simply depends on majority points and not on the surface we draw of something

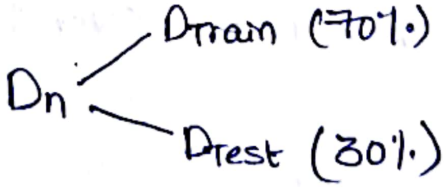
→ where as K=5 It is well-fitted fn.

Q.13 Need for Cross Validation :-

→ AS we see depending upon the K-Values we are Overfitting, underfitting fn's.

→ To determine 'K' or which K value will be suitable one idea is:

Consider Dn which is whole data set split this DataSet into Dtrain (70%.) Dtest (30%.)

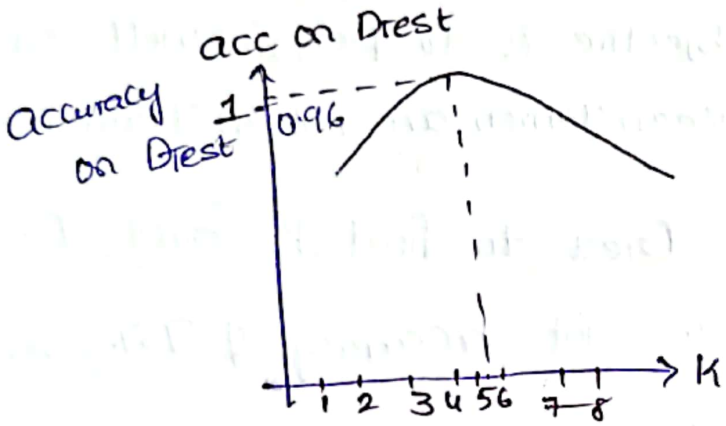


Now using Dtrain data by getting the accuracy on Dtest for different K values.

	Train	accuracy on Dtest
K=1	Dtrain	0.78
K=2	"	0.82
K=3	"	0.85

No of correctly classified pts / Total points

So we will check for different K values we will try to get the accuracy on Dtest with the help of Dtrain. The K value on which the accuracy which is typically higher will be considered



let K=6 gives me the best accuracy on Dtest when using Dtrain as Training data

we can say as

Using Dtrain & 6-NN on Amazon Reviews dataset

* I can get 96%. (if accuracy is 0.96)

* But there is a small problem.

as in General our objective is if we have the Dataset Dn the whole dataset Dn can be divided into two subsets Dtrain & Dtest using Dtrain & Dtest we tends to create an algorithm or function

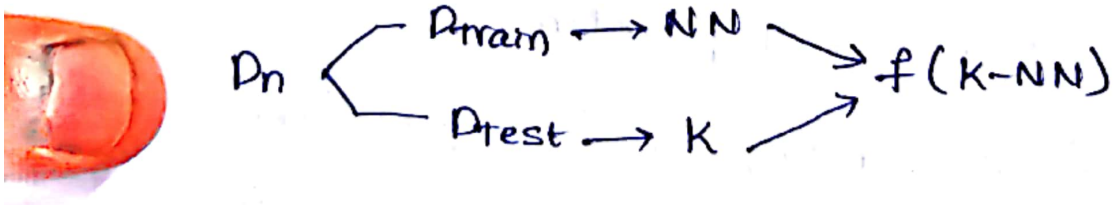
and for that function we try to get y_q by giving x_q. That means we try to apply the algorithm on future or unseen points. which is called "generalization."

Therefore as our objective is to perform well on unseen pts. (Unseen: which are not in Train & Test

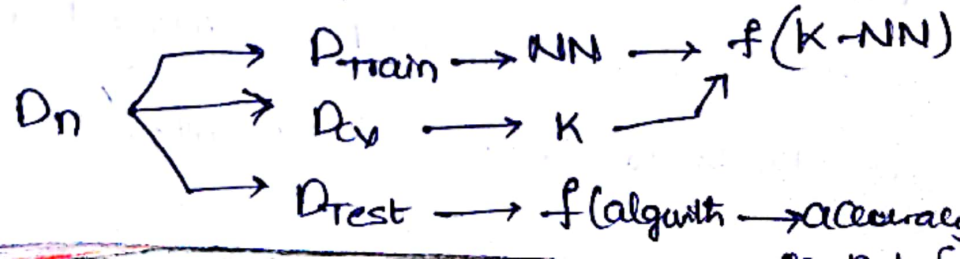
and as we used D_{test} to find K and D_{train} to find NN as if we got accuracy of 96% on D_{test} by using "K"-NN we cannot say that my accuracy on future data set or x_q would also be 96% by using Sam K.

TO Overcome This there IS Concept Called Cross-Validation (CV)

as in General Case we used



In Cross Validation



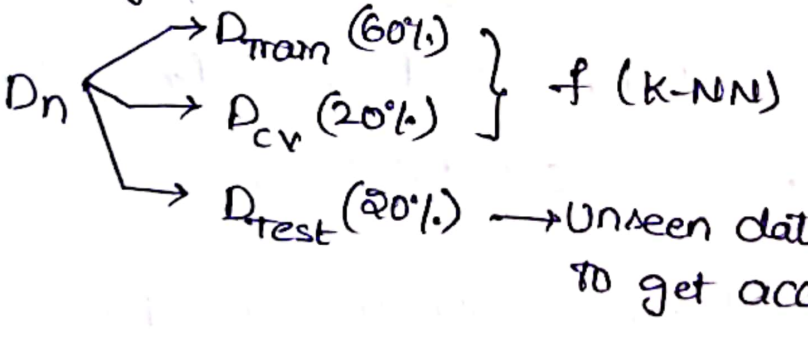
→ Now if we evaluate each point x_i in D_{test} as x_q and accuracy 93%. (as on unseen data) then we can say that 6-NN (K-NN) has an accuracy of 93% on unseen data, and this is called generalization accuracy.

and 100 - 93 = 7% is the generalization error on unseen data.

2.14 K'-fold Cross Validation:-

K-fold cross validation is where we combine D_{train} and D_{cv}

→ Generally we use whole data set D_n

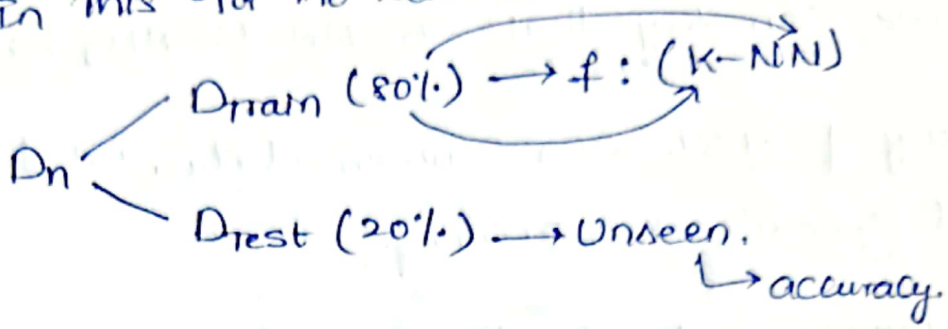


→ But in machine one of the key point is more the training data the better is your algorithms.

→ In K-fold cross validation we combine D_{train} and D_{cv} and total 80% data will be used to compute NN

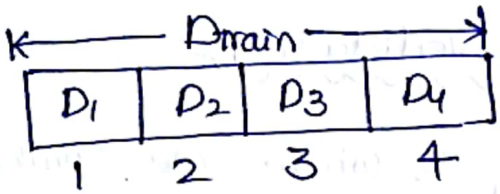
→ The K in k'-fold CV and K-NN in KNN both are different

Step 1:- So in this for the Dataset Dn



In Step 2:

we will try to break the whole Dtrain into 4 equal sized parts



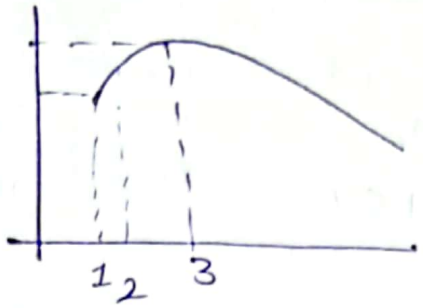
Step 3:-

we will find the accuracy on CV dataset.

as we only divided Dn into Dtrain & Dtest Then how could we get CV dataset?

So as we Divided Dtrain into 4 equal parts for different K values we will find the avg of accuracies and which ever 'k' gives the highest avg accuracy value that K needs to be considered

→ let us consider our accuracy on data set is like



		Train	CV	acc on CV	
4 times	K=1	$D_1 D_2 D_3$	D_4	a_4	avg $(a_1, a_2, a_3, a_4) = a_{K=1}$
	K=1	$D_1 D_2 D_4$	D_3	a_3	
	K=1	$D_1 D_3 D_4$	D_2	a_2	
	K=1	$D_2 D_3 D_4$	D_1	a_1	
Ily we find for K=2	K=2	$D_1 D_2 D_3$	D_4	a_4'	avg $a_{K=2}$
	K=2	.	.	.	
	K=2	.	.	.	
	K=2	.	.	.	

let the avg value we got is higher at K=3

→ as we are repeating for every K value four times so in K-fold K'= 4 here. So it is 4 fold C.V.

→ let K=3 has highest accuracy so we decided K=3 as best classifier so we get function as

f: 3-NN

→ So f: 3-NN → D_{Train}

↑

avg

accuracy on

4-fold CV

→ So if we given a review x_q we will use whole of Training data to compute NN and we use 3 NN to decide Class labels.

$x_q \xrightarrow{NN} D_{Train}$

→ 3 nearest neighbours

→ To calculate accuracy of the model.

we use D_{Test} →

D_{Test} → Accuracy of 3-NN on D_{Test}

let accuracy we got is 93%. on unseen data

Note :-

So avg-accuracy and accuracy of model is dif

By using avg accuracy we are find best k values depending upon which ever avg-accuracy value is hi and will be calculated with D_{Train}.

→ where as accuracy of the model will be calculated by Dtest.

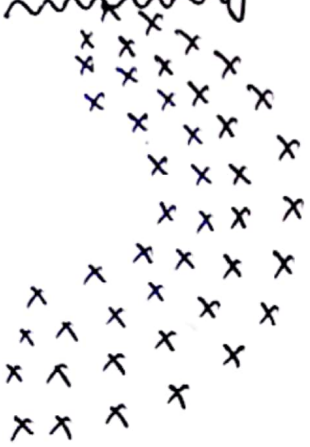
→ Note:-

Mostly used k'-fold CV is 10-fold CV. typically.

→ Note:-

Time it takes to compute the Optimal/best K in KNN increases by k' times if we ~~are~~ use k'-fold CV.

2.15 Visualizing, Train, CV & test datasets:-



when we are considering the whole dataset and we are randomly sampling the data set into