Date _____ Day _____

Validation data is used to evaluate the loss of a function "h" that is determined using the learning on the training data-set. If the loss on validation data is high for a given h, the hypothesis or model needs to be changed.

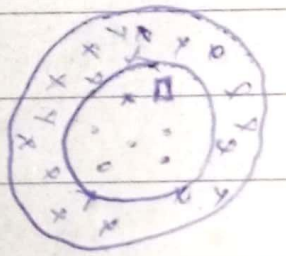
↳ helps improve the model by selecting the best configuration without overfitting.

↳ used to tune hypes parameters (like value of K in K-NN algo)

"K - NEAREST NEIGHBOUR ALGORITHM"

Space & Time complexity $O(dn)$ → d: no. of feature (dimension), n: no. of data points

↳ used to assign labels to unknown data points



	x_1	x_2	class	Distance
training data	1	5	0	2
	0	8	0	6
	0	6	1	4
	1	2	1	1

→ nearest

Test (1, 3) → calculated distances

↳ we need to identify the class

Date _____ Day _____

K=1 means only 1 nearest neighbour check kerye

=2 " two " " " " .

$$1 < K < N$$

↳ Hyper parameter

↳ For binary classification K = odd

↳ For multiple classification K = even

→ Complexity of prediction increases with size of training data.

→ Two types of distance Metric

- Euclidean distance = $\sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$
- Manhattan distance = $|x_2-x_1| + |y_2-y_1|$
- Minkowski distance = $\left[(|x_2-x_1|)^p + (|y_2-y_1|)^p\right]^{1/p}$

p = 1

↳ if p=1 → manhattan dist

p=2 so eucl dean dist

Date _____ Day _____

→ Miss match in the values of data
This can be solved thru normalization.

- Choice of K:
  - ▷ K = 1
    - ↳ if K=1, the model makes decision based on just one data point. If that point is incorrect, unusual or just slightly different, the entire prediction changes. This leads to overfitting
  - ▷ K = n
    - ↳ The algo assigns the majority class of the entire dataset
    - ↳ It ignores pattern & predicts the "most frequent" class
    - ↳ leads to underfitting
      - ↳ where model is too general & performs poorly on both test / training data.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n(\Sigma x^2) - (\Sigma x)^2][n(\Sigma y^2) - \ldots}}$$

Date -

8. As 3 is smaller than our best distance i.e we dont visit (2,3)
9. Now checking for 9,6
10. Ab 6,7 & 7,2 ka ⊥ draw krenge we get 1 (6-7) (x2 x level par hai to x waalo ke minus kra)
11. as 1 < best distance we go to 9,6
12. (6,7) & (9,6) d = 3.16 (donot update current best)
13. (6,7) & (8,1) d = 6 ...

So nearest is 4,7 ; Search bcm O(log n)

- Curse of Dimensionality:
  - Agr no. of dimensions zyada honge to data points dhoor dhoor honge.
  - Data points ko qareeb rakhne k liye we will need huge amount of data

Solution:
1. Increase dataset
2. Reduce dimensions

→ KNN carries out prediction about the test point assuming we have data points near to the test point that are similar to test point.
- As we donot have neighbours in high dimensionality space, KNN bcms vulnerable.

KNN → pg # 3c
28

Date _____ Day _____

→ All points tend to be almost equally far apart, making the concept of "nearest" unreliable.

▷ Parametric Algorithms :

↳ Assumes a fixed no. of parameters

↳ No matter how much data you throw it won't change it's mind about how many parameter it needs.

↳ Like in linear regression

$b0 + b1 * x1 + b2 * x2 = 0$

Coefficient → ek fixed functional form hai.

↳ The values for these is learned thru data

▷ Non Parametric Algorithms :

↳ No fixed parameters, model structure

Date _____ Day _____

Confusion Matrix

→ In skewed dataset we can't even tolerate 1% error rate.

→ Therefore to evaluate performance we use Confusion matrix

→ Details of 100 patients → 1% people have TB, 99% people have no TB

Actual \ Predicted	True	False
True	TP	FN → Type II error
False	FP → Type I error	TN

all diagonal values are TP

TP = True +ve = Correctly identified +ve

TN = True -ve = " " -ve

FN = False -ve = incorrectly identified +ve (Actual True hai lekin humne False predict kiya hai)

FP = False +ve = " " +ve (Actual False hai lekin humne True predict kiya hai)

• Precision Vs Recall

→ Precision tells us how many of the correctly predicted cases actually turned out to be +ve.

→ This would determine whether our model is reliable or not.

↳ Tells us accuracy of +ves.

in Email Spam detection.
→ Out of all emails you marked as SPAM, how many were actually SPAM.
→ If you marked 10 as Spam & 8 are actually SPAM then 80% precision

Page No.

For this to be max i.e = 1
then FP must be = 0

Date _____ Day _____

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{correctly identified +ves}}{\text{Total +ve Predicted}}$$

→ Recall tells us how many of the actual +ve cases we were able to predict correctly

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Correctly identified +ves}}{\text{Total Actual +ves}}$$

Out of all actual spam emails, how many did you correctly predicted?

if 20 spam emails were there & we correctly detected 8 then recall is 40%.

$$\text{F1-Score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$$

→ it is max when precision = recall

Micro avege =

Date _____ Day _____

Doctor testing for a disease

Positive → has ~~dis~~ TB

Neg → Donot have TB

Sick person diagnosed as sick

Healthy person diagnosed as sick

	Has TB	Dont have TB
predicted TB (True)	TP	FP
donot predicted TB (False)	FN	TN

A sick person dignosed healthy

A ~~sick person~~ healthy person diagnosed as healthy

Reall is usefull matric in cases where False Neg is higher concern than FP.

is important in medical cases. Q k agr ek sick ko healthy diagnosed krdia to masla hojayega.

Imbalance data:

Where one class has significantly more sample than the other in a classification problem.

Eg. a fraud detection dataset → 95% loan are normal
→ 5% are fraudulent

Date _____ Day _____

Predicted \ Actual	Dog	Cat	Rabbit
Dog	23 TP	12 FP	7 FP
Cat	11	29 TP	13
Rabbit	4	10	24 TP

Predicted \ Actual	Dog	Cat & Rabbit
Dog	TP (23)	FP 12+7 (19)
Cat & Rabbit	FN 11+4 (14)	TN 29+13+10+24 76

	Actual Cat	Actual Dog & Rabbit
Cat	TP 29	FP 11+13 24
Dog & Rabbit	FN 12+10 22	TN 23+7+4+24 58

Predicted \ Actual	Rabbit	Dog & Cat
Rabbit	TP 24	FP 4+10 (14)
Dog & Cat	FN 7+13 20	TN 23+7+4+29 75

	TP	TN	FP	FN
Dog	23	76	19	15
Cat	29	58	24	22
Rabbit	24	75	14	20

Page No. _____

Macro → along along
Micro → ek saath

A	50	4	2
B	5	45	5
C	3	6	40

Date _____ Day _____

Accuracy =

Predicted \ Actual		A	B	C
	A	50	4	2
	B	5	45	5
	C	3	6	40

FP → row
FN → col

Predicted \ Actual		A	B ∪ C
	A	TP 50	FP 4+2 = 6
	B ∪ C	FN 5+3 = 8	TN 45+5 = 96 (+6 +40)

Predicted \ Actual		B	A ∪ C
	B	TP 45	FP 5+5 = 10
	A ∪ C	FN 4+6 = 10	TN 50+2 = 95 (+3+40)

Predicted \ Actual		C	A ∪ B
	C	TP 40	FP 3+6 = 9
	A ∪ B	FN 2+5 = 7	TN 50+4 = 104 (+5+45)

$$\text{Accuracy} = \frac{TP_A + TP_B + TP_C}{\text{Total Samples}} =$$

Formulas For Distance

$|x_2 - x_1| + |y_2 - y_1|$ → manhattan

$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ → euclidean

$1 < K < N$

Date: ____________

K-Nearest Neighbour (KNN)

↳ can be used for classification as well as regression

training data:

x_1	x_2	class	Distance
1	5	0	2
0	8	0	6
0	6	1	4
1	2	1	1

Test : (1,3)

distance formula se tamam training data se distance nikaal liya

we need to find class of this test point

if K=1 means we need to find one nearest neighbour.

↳ we choose odd values (q k agr ek se 1 class ka label agya doosre se 0 to kisi ek ko choose nhi kr sakte.

"K" ~~can~~ ≠ no. of data points in training data.

Height	Weight	Label	Distance
170	70	Normal	16
160	60	Normal	36
180	80	Over	4
175	75	over	6
182	82	overweight	8
169	66	normal	21

Test (178,78)

K=3

label for Test ie 170

→ [redacted]

Date:

If K = 1, then of K ka model overfit hojayega
iska sirf ek hi point ko learn kreg as
K=1. Training accuracy to achi hogi (ekn
testing accuracy achi nhi aayegi.
↳ overlearn

If K = N, then model is underfit.
Pattern nhi learn kr payega.
Testing & Traing both accuracy is low.
↳ no learning

→ if Binary classification then choose odd value

→ if Multiple a the even.

- Document A (Sports): "Team wins championship match."
- Doc B (Politics): "Election results were announced yesterday."
- New Doc (?): "The match results were announced"

we need to identify label for this