academics | INSAID .co

DEEPESH.

K Nearest Neighbor

K, dist

Wine dataset

√n : CV :

distance

03 Aug 19 theory + concepts

04 Aug 19 practical

TIME TABLE

1	Sat 27 July 19	PCA	
	Sun 28 July 19	PCA + practical	
2	Sat 03 Aug 19	KNN	KNN
	Sun 04 Aug 19	KNN	K means
3	Sat 10 Aug 19	K means	K means
	Sun 11 Aug 2019		
4	Sat 17 Aug 2019		
	Sun 18 Aug 2019		
5	Sat 24 Aug 2019		
	Sun 25 Aug 2019		
6	Sat 31 Aug 2019		
	Sun 01 Sep 2019		

ML 2

2 → 1. Principal Component Analysis (PCA)

1½ 2. K Nearest Neighbors (KNN)

3. Naive Bayes

4. Support Vector machines (SVM)

½ → 5. K means (clustering)

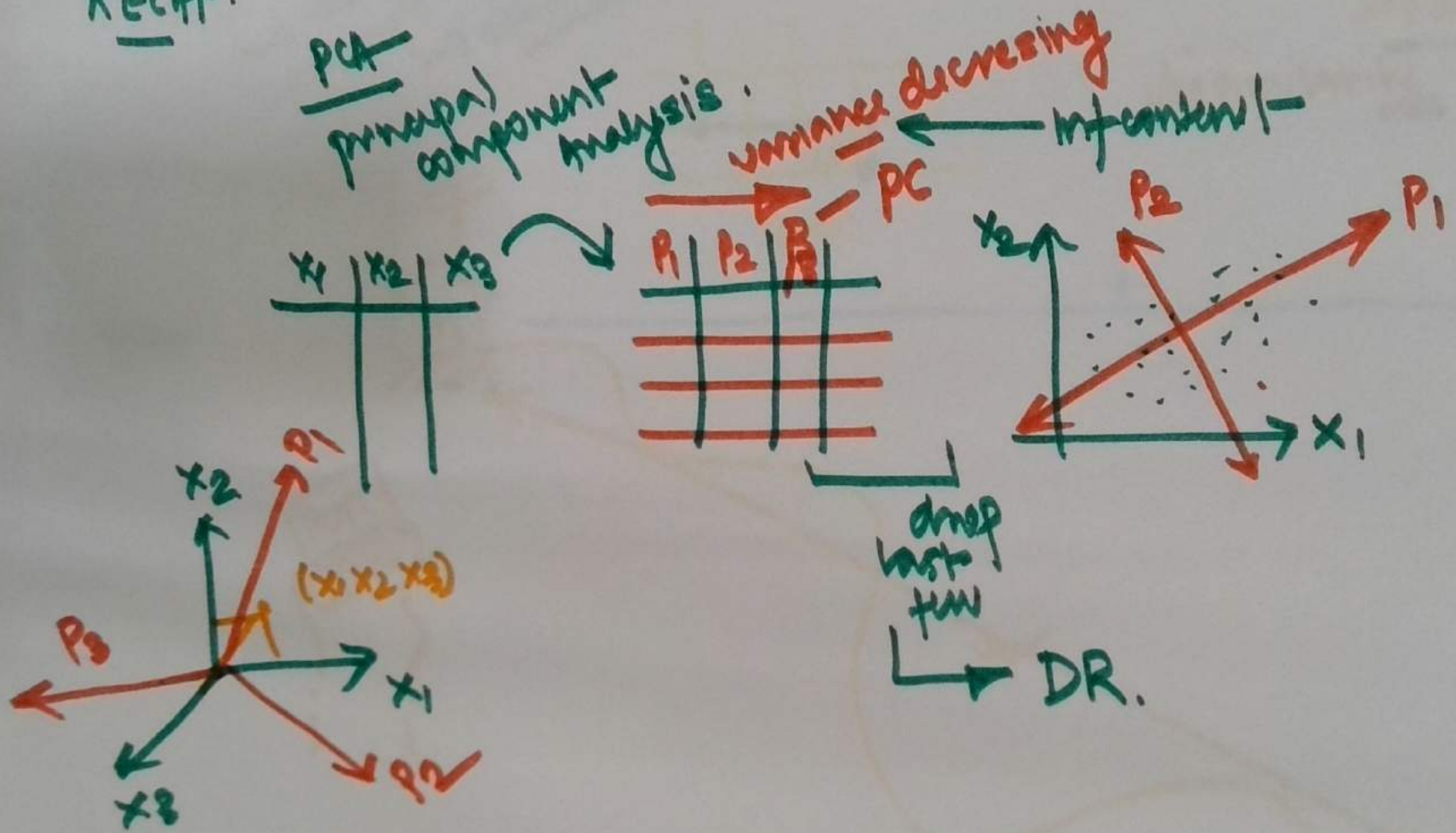
6. Time series Analysis

≥20% practical

RECAP.

PCA

principal component analysis.

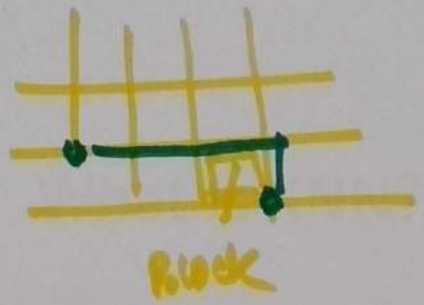


3

KNN → nearest → distance

"DISTANCE"

dissimilarity



Time t

15 mins

Y

2 song lengths away

X

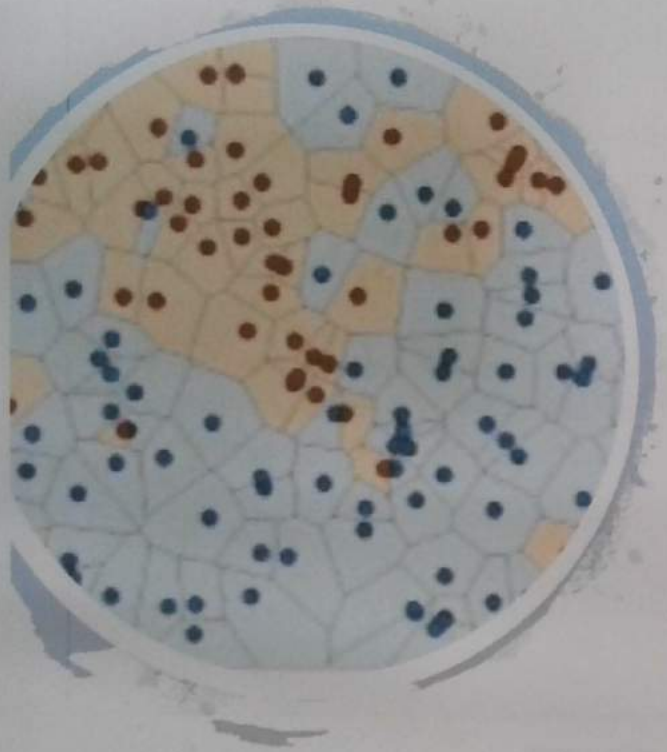
Item

location

X, Y

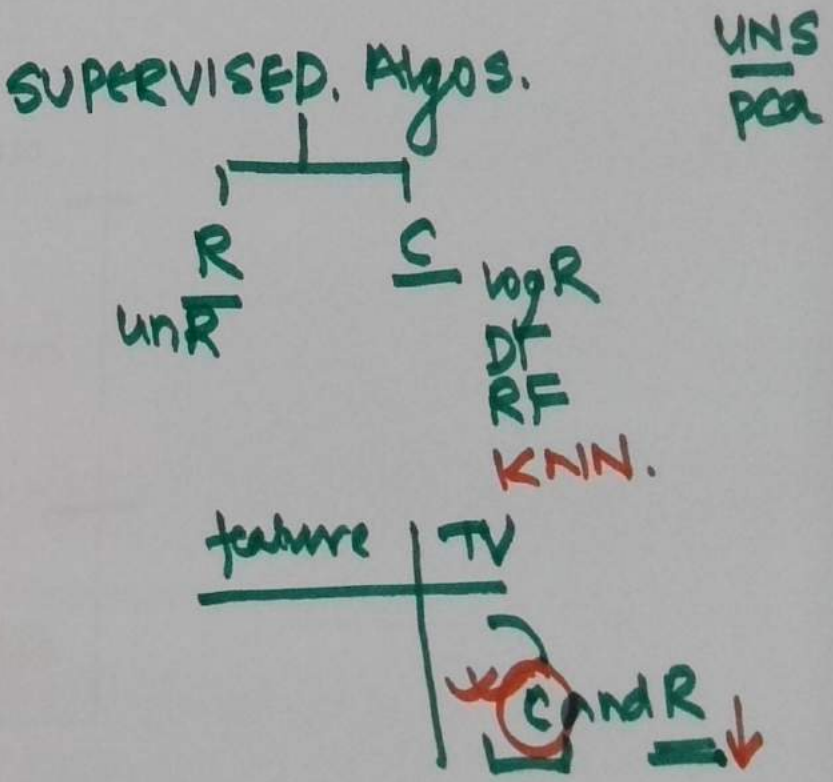
not lat/long

4



What is KNN?

- It is one of the simplest Supervised Machine Learning algorithm.
- K nearest neighbors stores all available cases .
- It classifies new cases based on a similarity measure(eg; distance).



5

INSAID

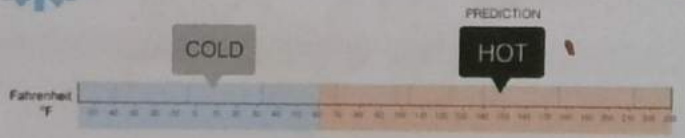
- It is used for both classification and regression predictive problems.
- However, it is more widely used in classification problems in the industry.

Regression

What is the temperature going to be tomorrow?

Classification

Will it be Cold or Hot tomorrow?



KNN is based on Feature similarity.

INSAID

International School of AI & Data Science

Algorithm.

CAT's → CAT.

DOGs

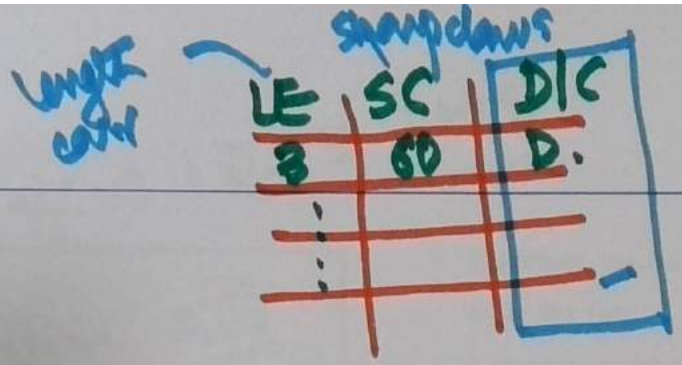
NN
SVM
DT, RF

Dissimilarity in pop

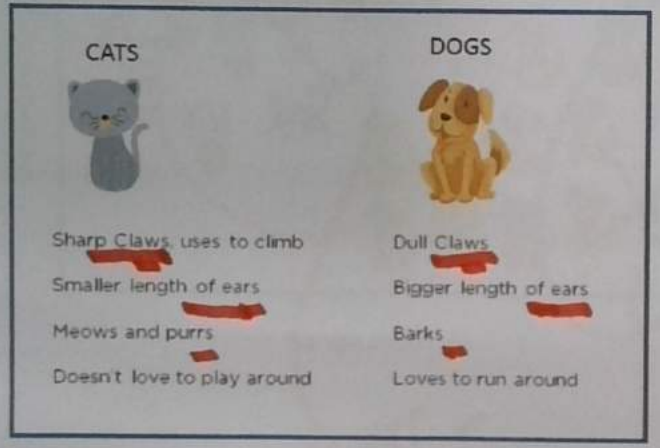
INSAID

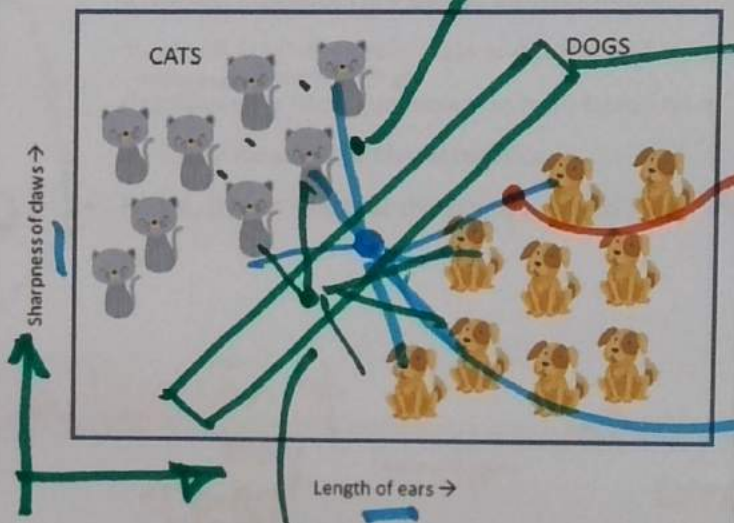
Lets consider an example:

8



INSAID





unknown animal ≡ CAT

unknow ≡ DOG

unknown ≡ CAT or DOG.

WHAT IF DISTANCE IS EQUAL?

Check K of nearest neigh.

5

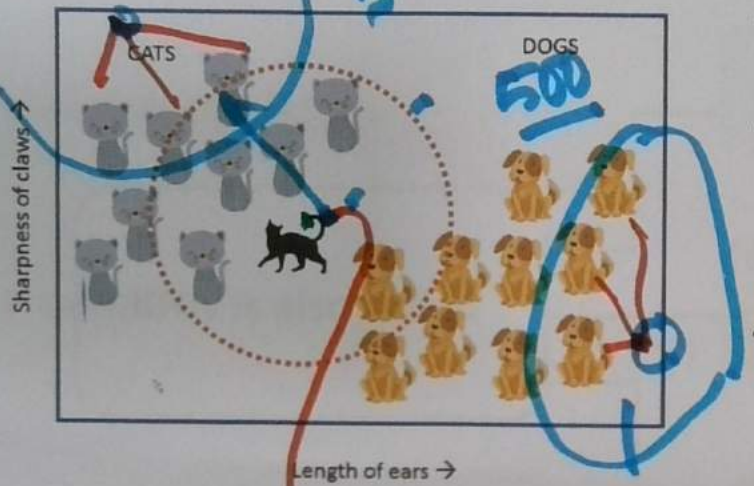
3/5 ≡ DOG

2/5 ≡ CAT

International School of AI & Data Science

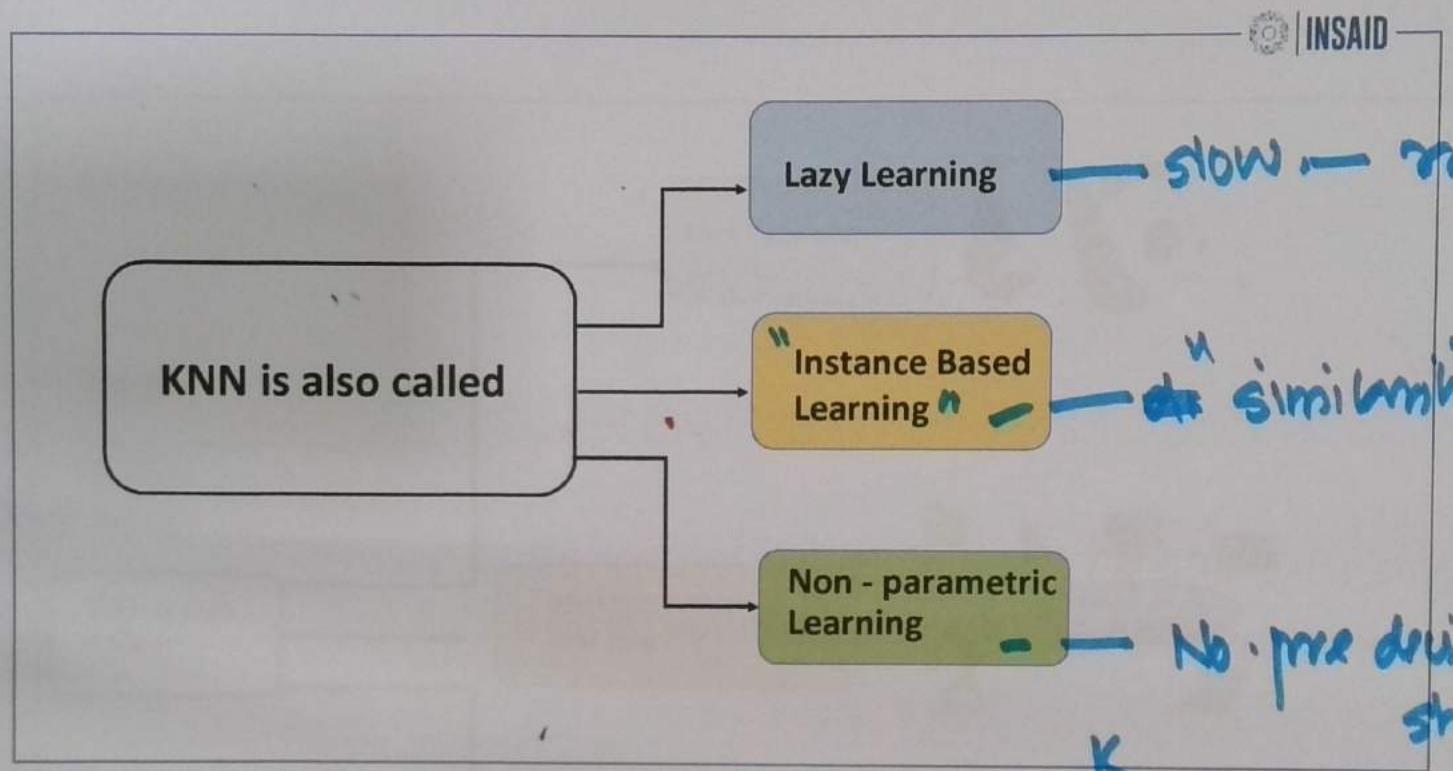
9

INSAID



- Therefore, the features of the strange animal are more like cats.
- K nearest neighbors is a simple algorithm that stores all available cases (cats and dogs) and classifies new cases (unknown animal) based on a similarity measure or characteristics.
- Hence, it is proved that the strange animal is a cat.

10



slow — ran algo independent for all data points

"similarity based"

No. pre decided structure

K dist.

LogR - para.
$y = \beta_0 + \beta_1 x$

DT - np.

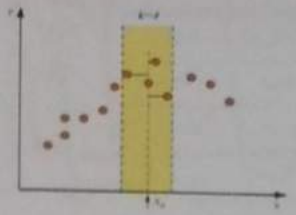
INSAID

KNN is also called

Lazy Learning

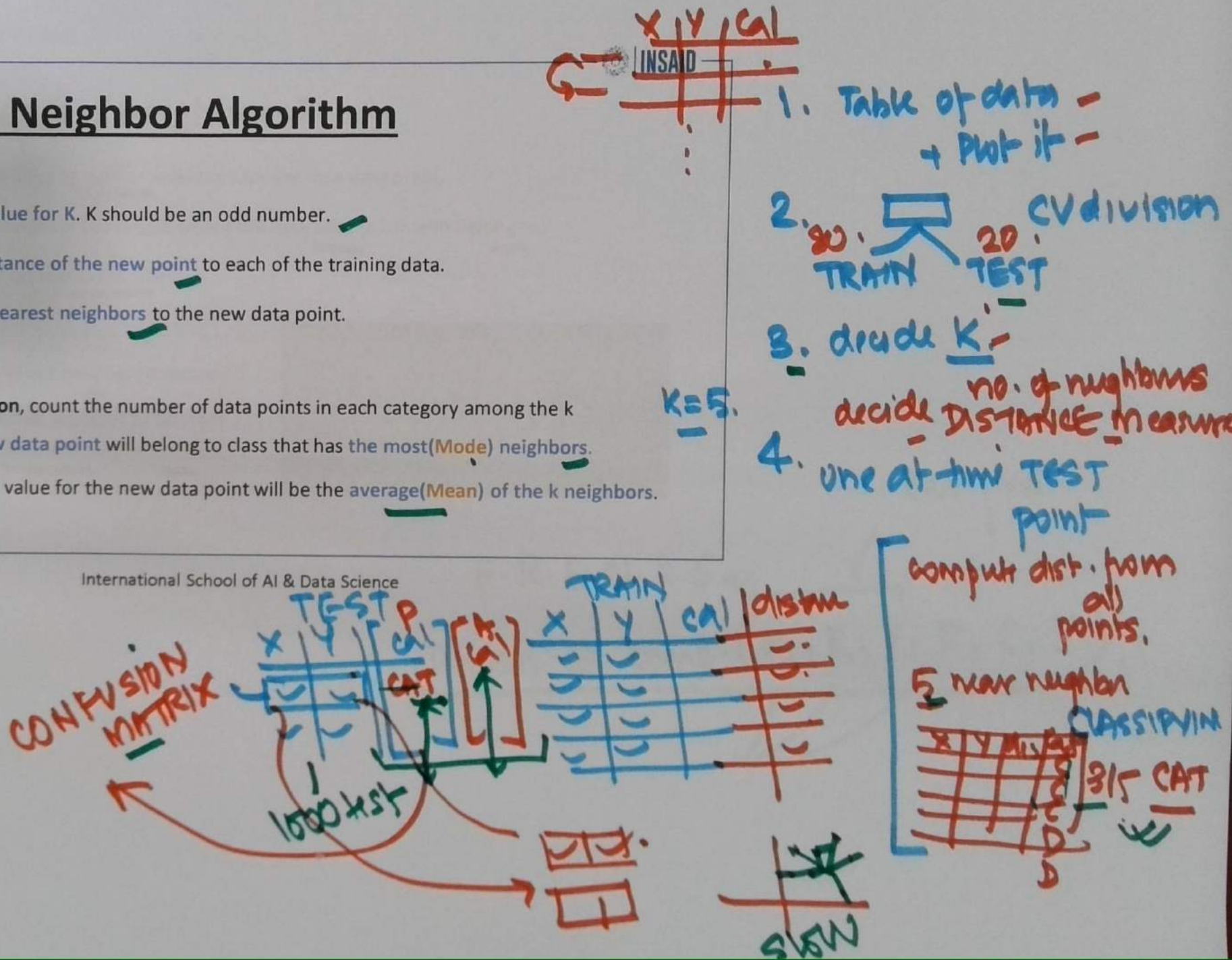
Instance Based Learning

Non - parametric Learning



K Nearest Neighbor Algorithm

- **Step 1**: Choose a value for K. K should be an odd number.
- **Step 2**: Find the distance of the new point to each of the training data.
- **Step 3**: Find the K nearest neighbors to the new data point.
- **Step 4**:
  - **For classification**, count the number of data points in each category among the k neighbors. New data point will belong to class that has the most(Mode) neighbors.
  - **For regression**, value for the new data point will be the average(Mean) of the k neighbors.



INSAID

What is K nearest neighbors?

- K is a number used to identify similar neighbors for the new data point.
- KNN takes K nearest neighbors to decide where the new data point with belong to.
- This decision is based on feature similarity.

For example

- We have Friend circle in our new neighborhood.
- We select 3 neighbors that we want to be close friends based on common thinking or hobbies.
- In this case K is 3

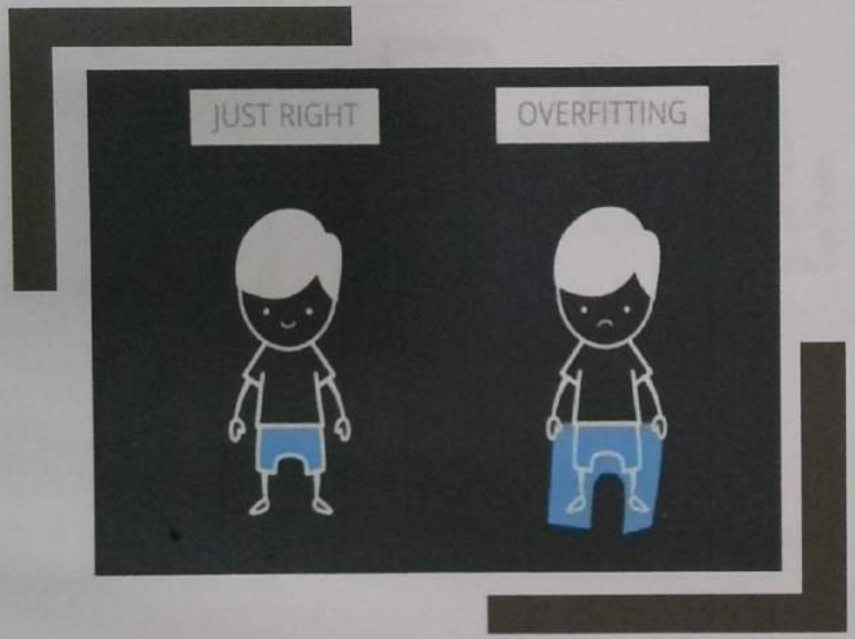
our pals

F.R.E.N.D.S

Person ≡ comb.(M, R, P, R, C, J)

INSAID

Overfitting in KNN



- When "k" is a **very small number- KNN can overfit.**
- It will classify just based on the **closest neighbors** instead of learning a good separating frontier between classes.

K

K=1

RED

GREEN

OVERFIT

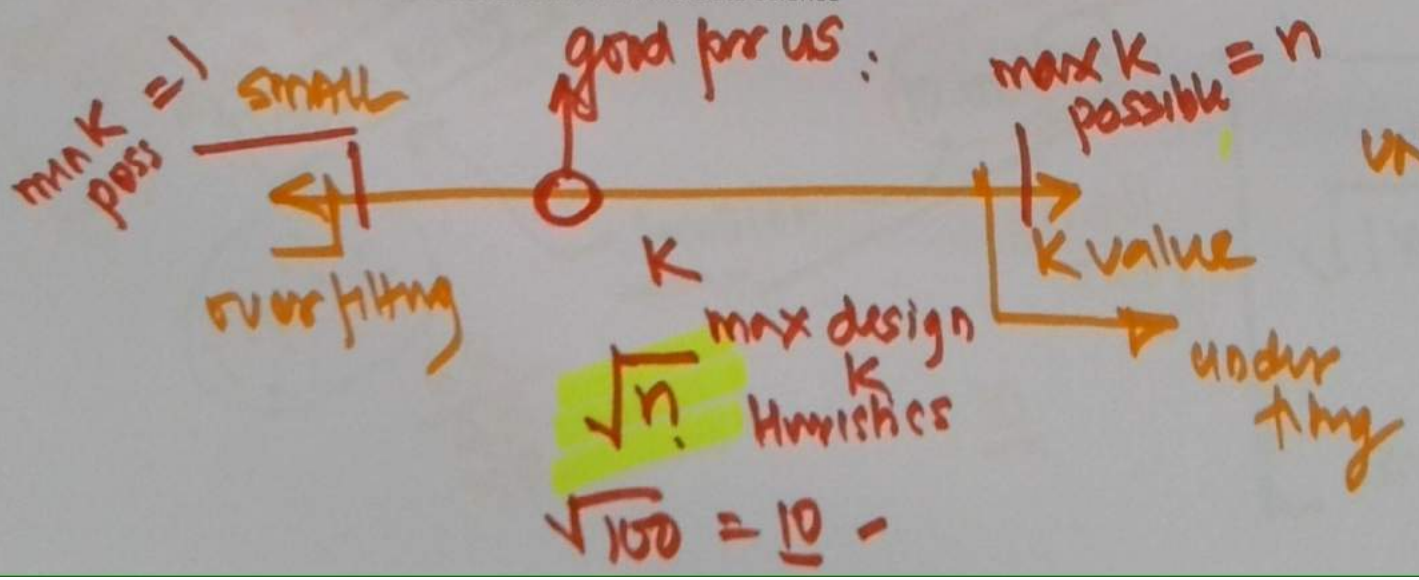
K=1

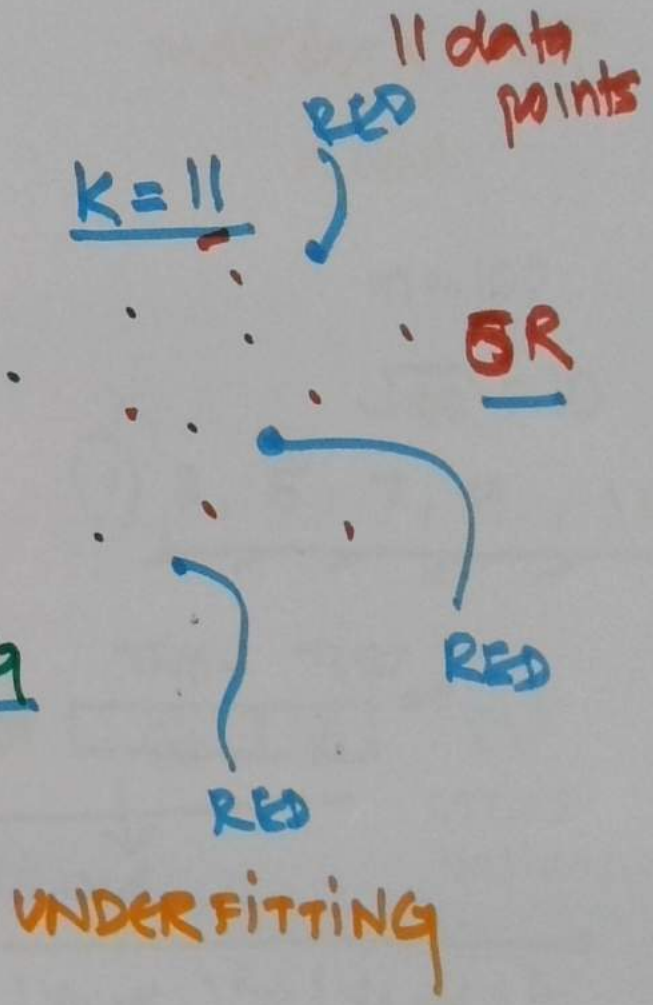
Small K → overfitting

INSAID

Underfitting in KNN

- If **k = n**, KNN will think every point belongs to the class that has more samples.
- If "k" is a **very big number** - KNN will underfit, in the limit.

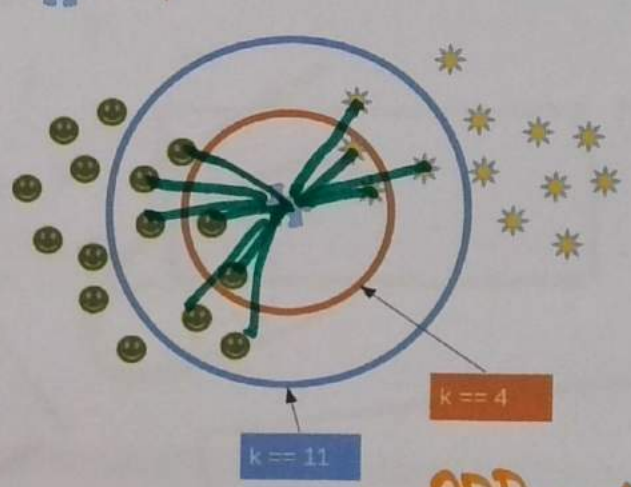




INSAID

How do we choose the factor K ?

- Sqrt(n), where n is the total number of datapoints.
- Odd value of k is selected to avoid any confusion between two classes off data.
- In python, we will see how to use cross validation to choose the factor K(Later).



max odd K = $\sqrt{n}$

K odd.

n = 100

$\sqrt{100}$ = 10

1 3, 5, 7, 9, 11

ODD

EASY to EXPLAIN.
SIMPLE
DECENT ACCURACY.

10000 points

1million

10 million.

SMALL TEST SET

6R
4B
5R

10

TRAIN TEST:

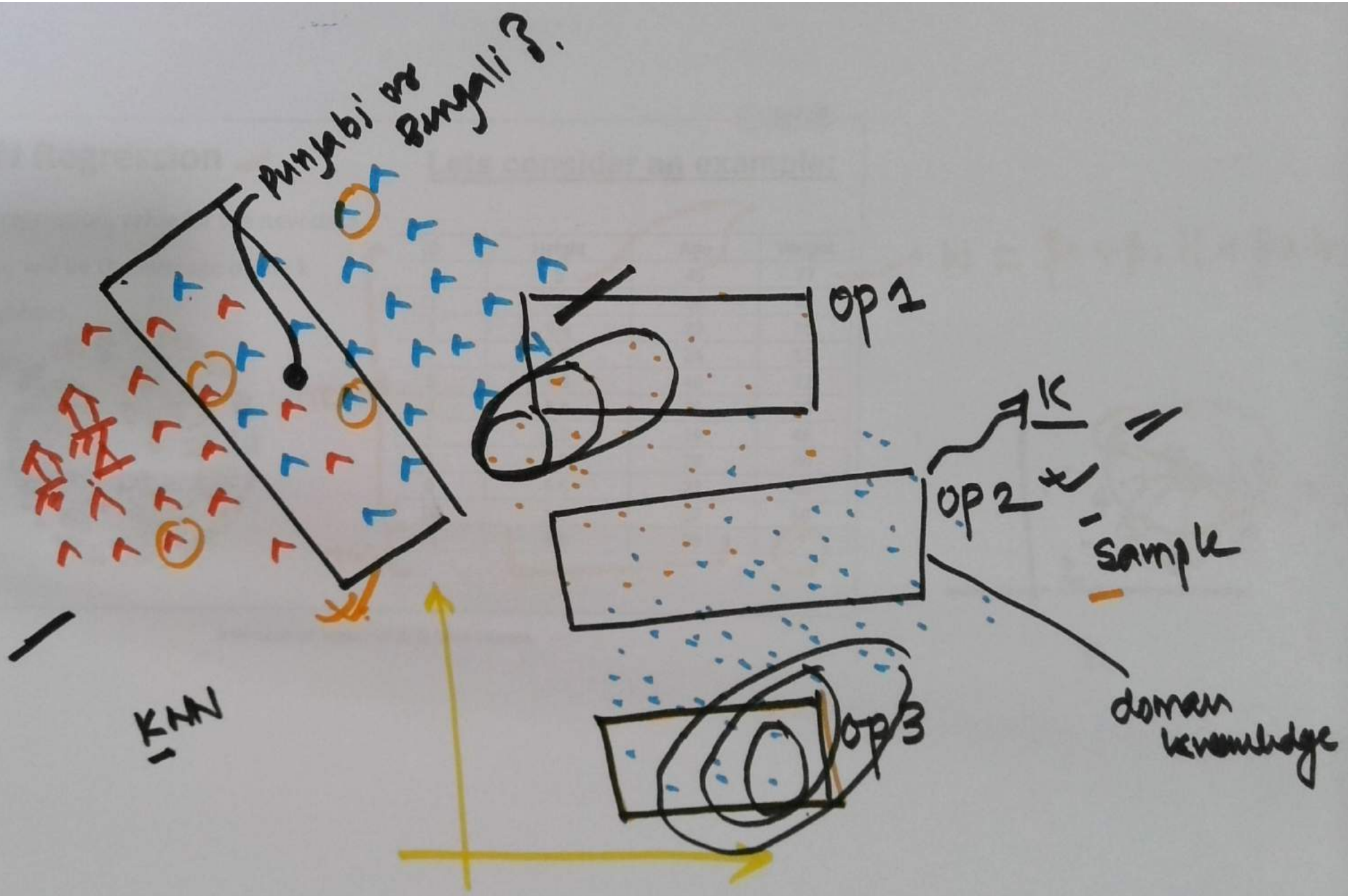
80 20

CV
cross validation

K

$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2 + (z_2-z_1)^2 \ldots}$

confusion. = =


Punjabi or Bangali?
op 1
op 2
sample
KNN
op 3
domain knowledge

INSAID

KNN Regression

Lets consider an example:

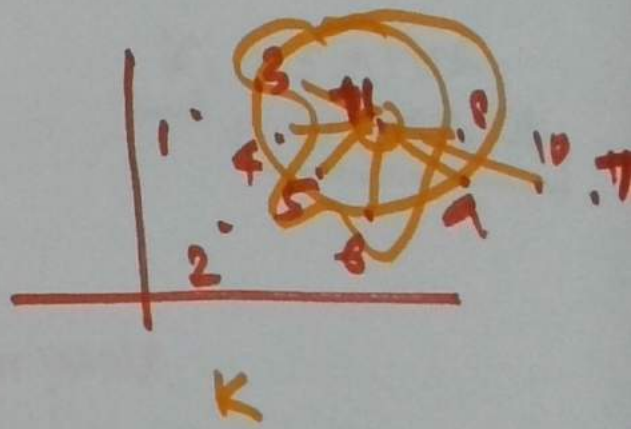
- For regression, value for the new data point will be the average of the k neighbors.

ID	Height	Age	Weight
1	5	45	77
2	5.11	26	47
3	5.6	30	55
4	5.9	34	59
5	4.8	40	72
6	5.8	36	60
7	5.3	19	40
8	5.8	28	60
9	5.5	23	45
10	5.6	32	58
11	5.5	38	?

TRAIN

TEST

$W = \beta_0 + \beta_1 H + \beta_2 A$



19

K=5

5 45

5.5 , 38

INSAID

ID	Height	Age	Weight	Distance
1	5	45	77	7.018
2	5.11	26	47	
3	5.6	30	55	
4	5.9	34	59	
5	4.8	40	72	
6	5.8	36	60	
7	5.3	19	40	
8	5.8	28	60	
9	5.5	23	45	
10	5.6	32	58	
11	5.5	38	?	

TRAIN

TEST

12

13

14

distance $(5.5-5)^2 + (45-38)^2$

$= 7.018$

K: distance measure

average

52.8

5 nearest data points

Regression

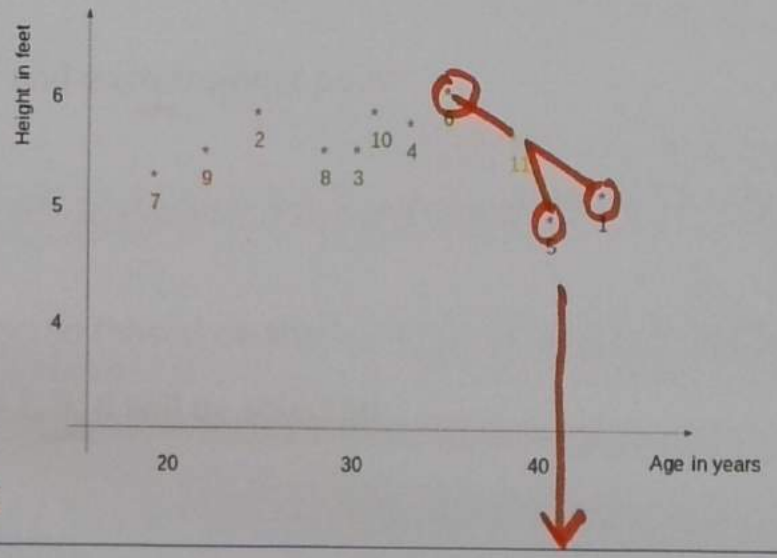
5NN

INSAID

Below is the plot of height versus age from the table:

K=3

- Y-axis represents the height of a person (in feet)
- X-axis represents the age (in years).
- The points are numbered according to the ID values.
- The yellow point (ID 11) is our test point.



calculate all 10 distances

↓

shortest 3

average ≡ REG.

1 2 3 4

≈ 3.57

Step 1

Distance between the new point and each training point is calculated.

Step 2

The closest k data points are selected (based on the distance). In this example, points 1, 5, 6 will be selected if value of k is 3

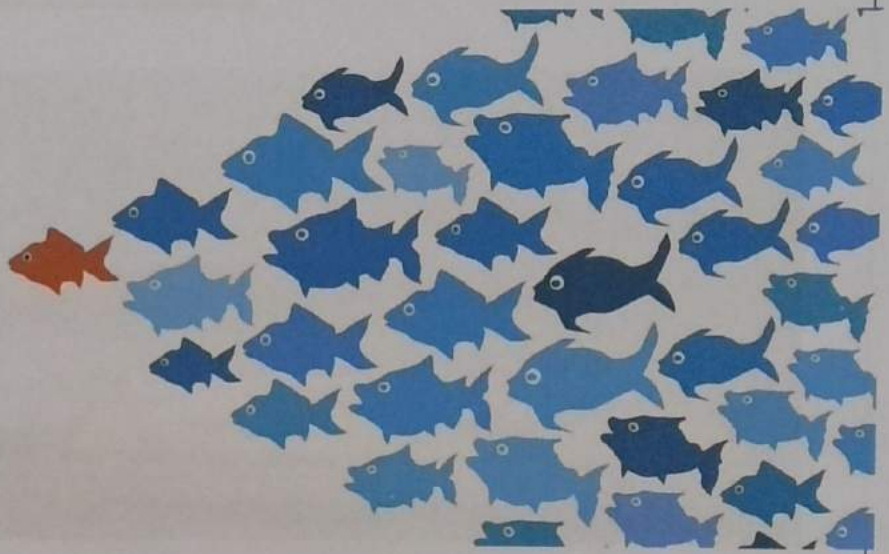
Step 3

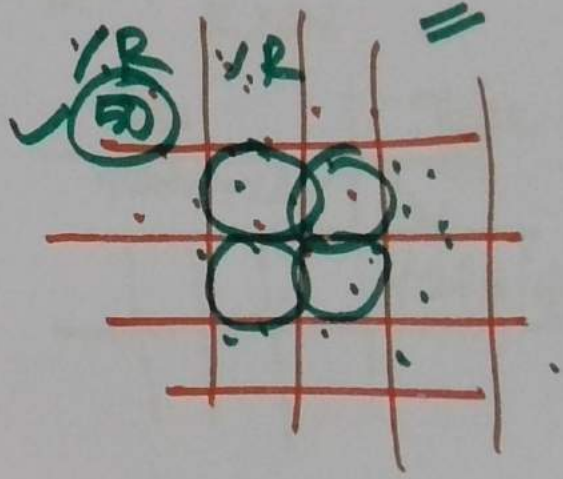
The average of these data points is the final prediction for the new point. Here, we have weight of ID11 = (77+72+60)/3 = 69.66 kg.

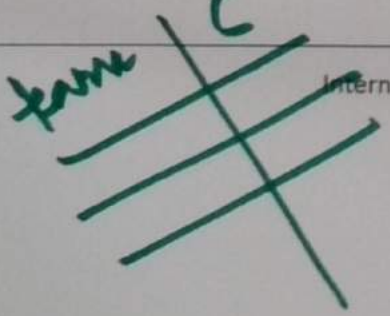
INSAID

KNN Classification

- For classification, count the number of data points in each category among the k neighbors.
- New data point will belong to class that has the most neighbors.



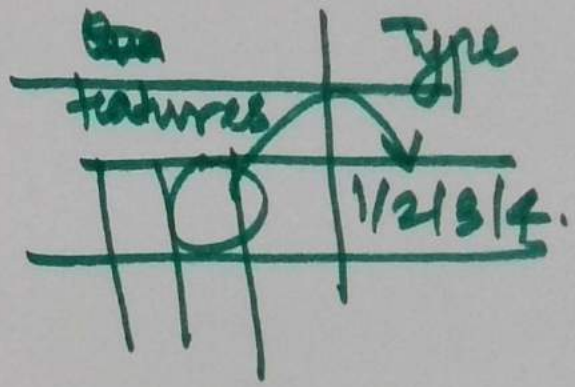




Wine Quality database – KNN

Lets consider an example:

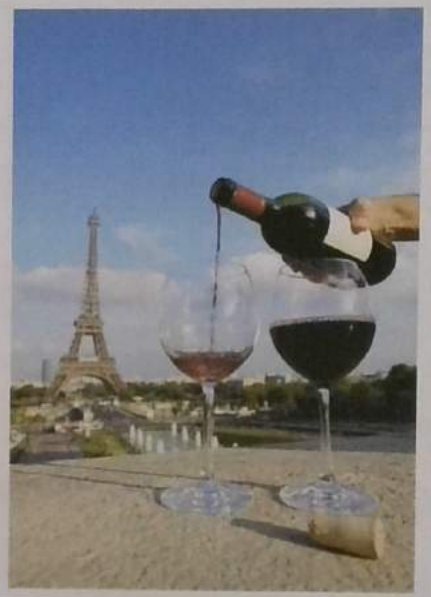
Wine Quality Detection with KNN Algorithm



INSAID

What if want to classify a new wine quality ?

- In this case , we would find the distance between this wine feature with all the wines.



INSAID

Wine Dataset

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

- **The dataset comprise of 12 features.**
- We classified wine on the basis of closeness **of different features.**

TV

Experiment K =

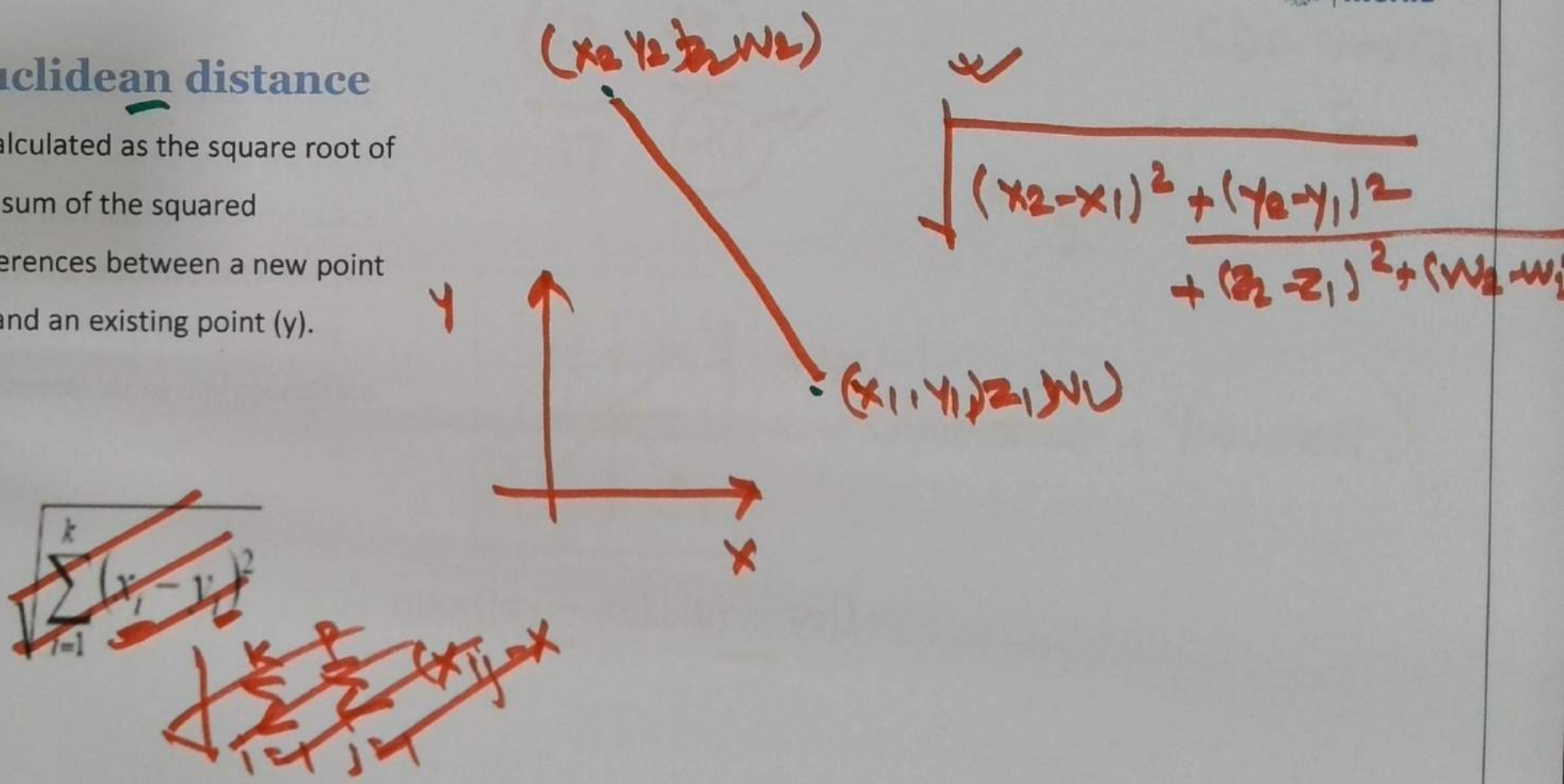
Accuracy

K

Distances =

Euclidean distance

is calculated as the square root of the sum of the squared differences between a new point (x) and an existing point (y).



(45, 5)

((62, 25)

17 20

INSAID

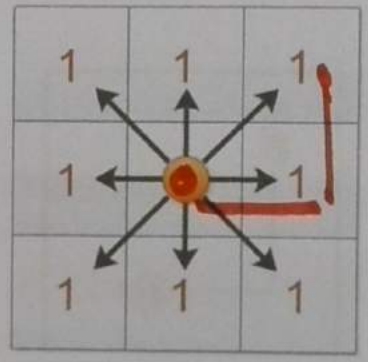
CD = max(2, 1)

= 2

2 1

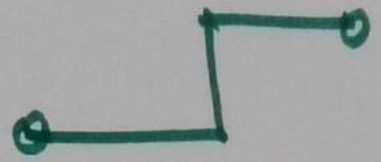
Chebyshev distance

maximum metric, or L_∞ metric is a metric defined on a vector space where the **distance** between two vectors is the greatest of their differences along any coordinate dimension.



max (Xmovement , Ymovement)

$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

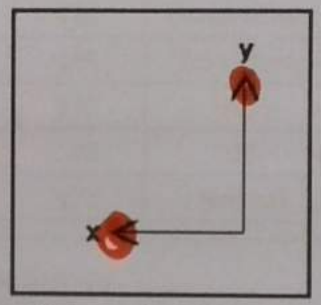


Manhattan distance

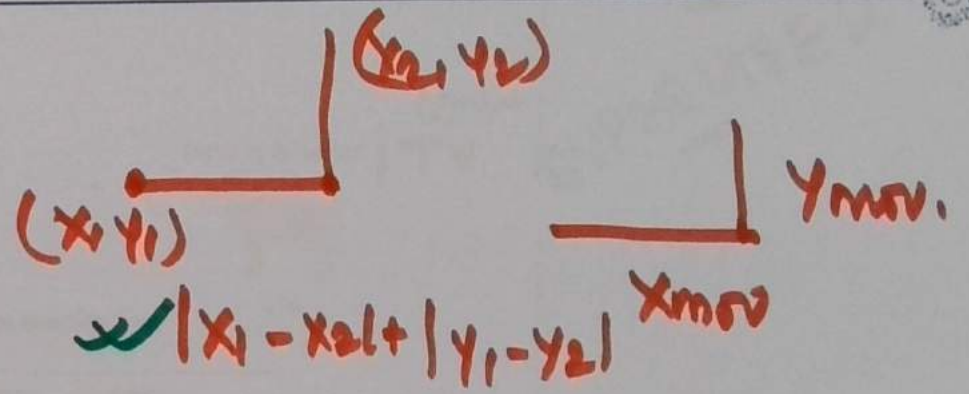
between two vectors (or points) a and b is defined as

$$D = \sum_{i=1}^{n} |x_i - y_i|$$

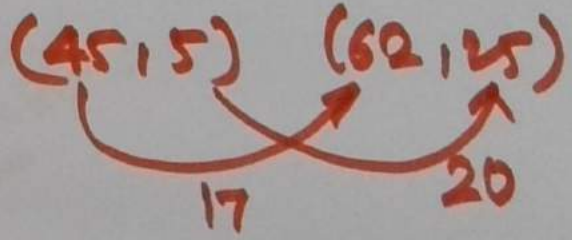
This is known as Manhattan distance because all paths from the bottom left to top right of this idealized city have the same distance:



Manhattan



Xmov + Ymov.



MD = 17 + 20 = 37

INSAID

When do we use KNN?

We can use K Nearest Neighbor when

Data is labelled

Dog

Dataset is small

Because KNN is a "lazy learner"

Dataset is noise free

Weight	Height	Class
51	182	Underweight
62	165	One - forty
69	176	Hello
64	173	23
65	172	Normal

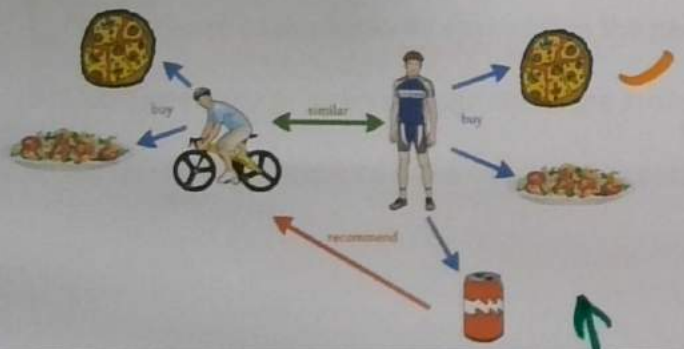
Noise

INSAID

Applications

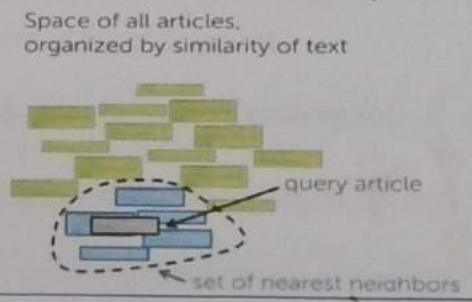
Recommender system

If you know a user likes a particular item, then you can recommend similar items for them.



Document Retrieval

We compute distances from query article to all other articles. Then we search for the articles with smallest distance to the query article.
They are called nearest neighbors.



INSAID

Advantages of K-nearest neighbors algorithm

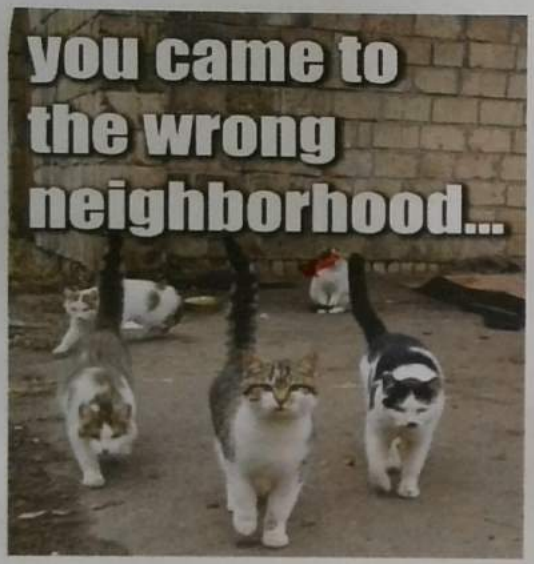
- It is simple to implement.
- It executes quickly for small training data sets.
- Performance asymptotically approaches the performance of the Bayes Classifier.
- Don't need any prior knowledge about the structure of data in the training set.
- No retraining is required if the new training pattern is added to the existing training set.

Naïve Bayes

BAYES CALASSIFICATION

INSAID

Limitations of K-nearest neighbors algorithm



- When the training set is large, it may **take a lot of space.**
- For every test data, the distance should be computed between test data and all the training data. Thus a **lot of time** may be needed for the testing.