

Session - 2

KNN - 2



## Agenda

- ① Recap
- ② BIAS- VARIANCE
- ③ HYPERPARAMETER TUNING
- ④ KNN For CATEGORICAL FEATURES
- ⑤ DISTANCE METRICS
- ⑥ KNN IMPUTATION

## Bias and Variance

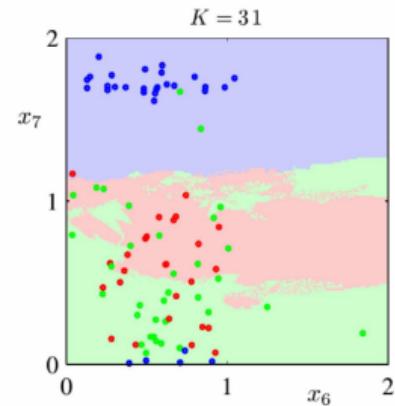
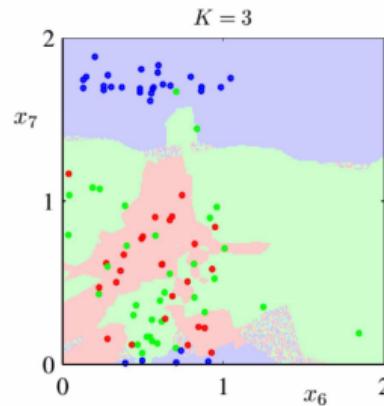
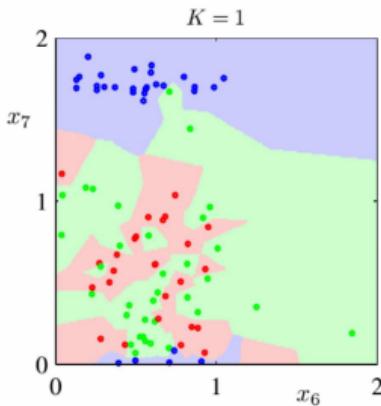
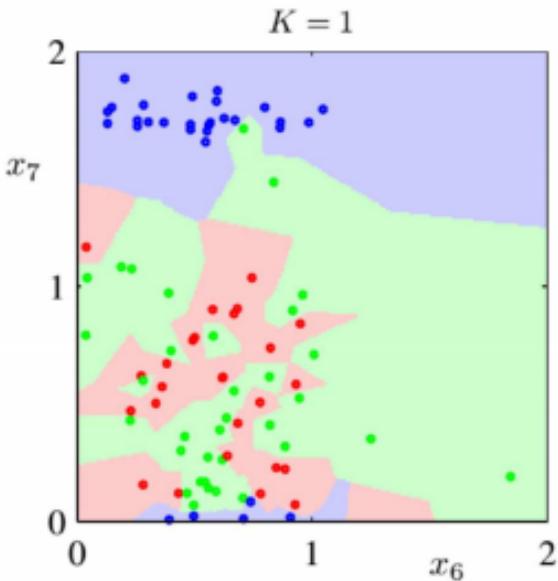
<https://datascience.stackexchange.com/questions/81866/why-does-the-overfitting-decreases-if-we-choose-k-to-be-large-in-k-nearest-neigh>

Bias: Errors made by my model on test/production

Data because of making very simple assumptions on my training data.

Variance: Errors made by my model on test/

production data because of making very complicated assumptions on my training data.



$$K = 100$$

$$P_1 - d_1$$

$$P_2 - d_2 \rightarrow$$

,

,

$$P_{100} - d_{100}$$

Sort

$$P_6 \rightarrow d_6$$

$$P_7 \rightarrow d_7$$

,

,

,

,

,

,

,

,

,

,

,

,

,

,

,

,

,

,

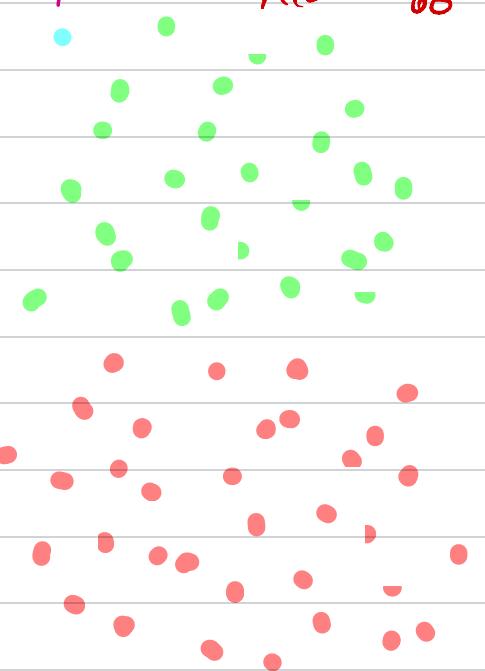
,

10<sup>1</sup>

/

Gr - u<sub>0</sub>

Red - 60

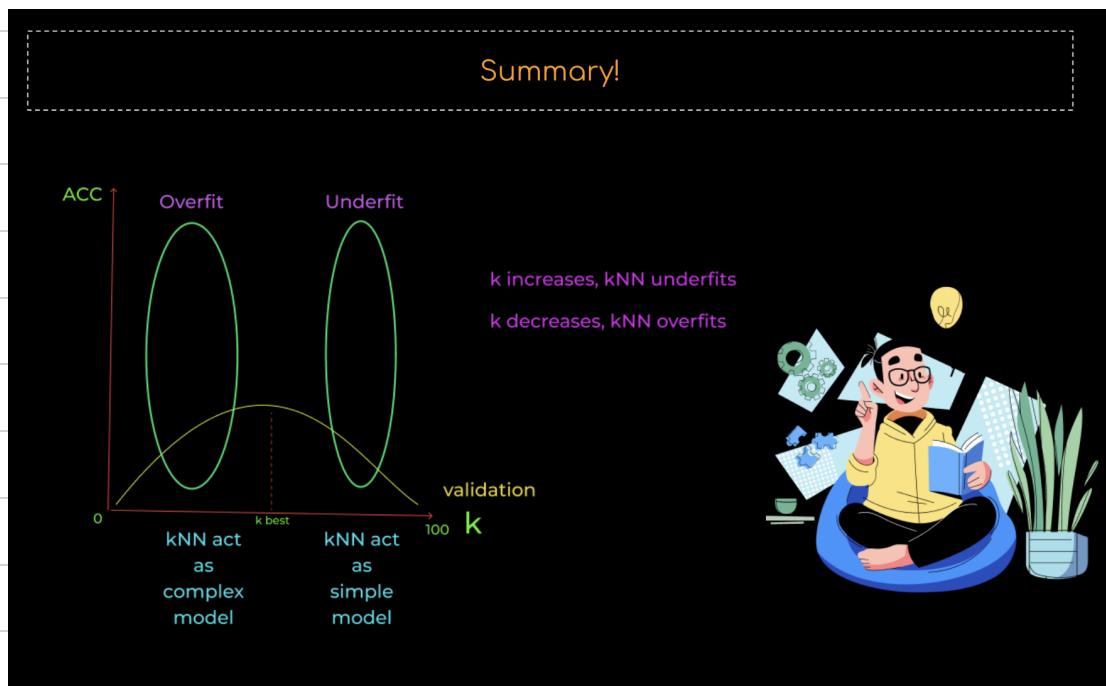
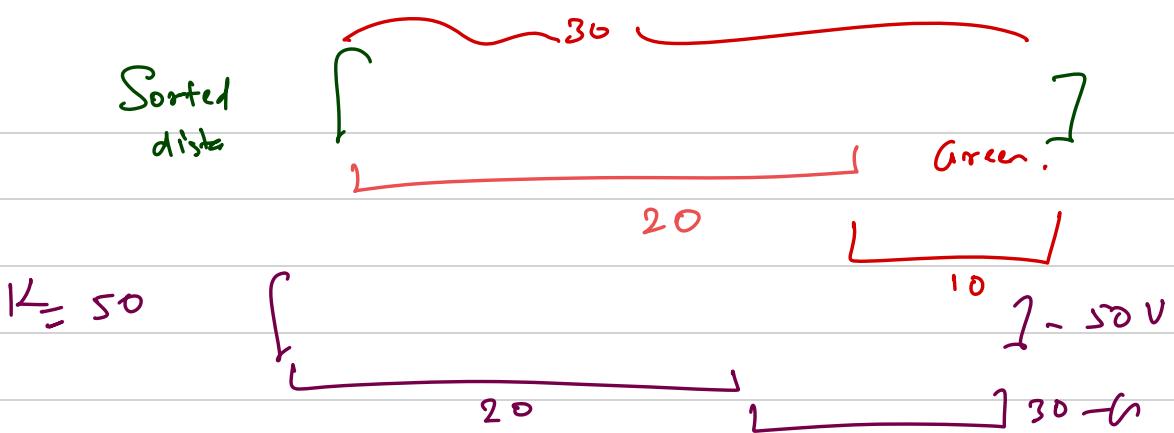


CA  
20

a  
n.

u<sub>0</sub> - B  
b

$$K = \underline{30}$$



## what to say if model has high variance and low bias ?

0 users have participated



A Model overfits

0%



PC  
2  
Pankaj Chaudhary  
1/1 ⚡ 97.10



AS  
1  
Aditya Shandilya  
1/1 ⚡ 97.36



AS  
3  
Ayush sachan  
1/1 ⚡ 97.00

B Model underfits

0%

[End Quiz Now](#)

k→ hyperparameter, then what data to use for hyperparameter tuning ?

4 users have participated



A validation

50%

B training

50%

C test

0%

D entire data

0%

[End Quiz Now](#)

Based on all quizzes from the session



SJ  
2  
SHASHANK JHA  
2/2 + 191.37



AS  
1  
Aditya Shandilya  
2/2 + 193.70



NP  
3  
Nachiket Pawar  
2/2 + 190.20

4	AD	Aditya Shandilya	2/2 ⚡ 188.13
5	SA	Souvik Adhikary	2/2 ⚡ 187.53
6	PS	Perisetta Pavan Kalyan	2/2 ⚡ 186.86
7	KP	Karthik	2/2 ⚡ 182.26
8	GV	G Vibu Vignesh	2/2 ⚡ 172.55
9	OP	OM PRAKASH S	2/2 ⚡ 166.96
10	ST	Sakilur Rahaman Thander	2/2 ⚡ 165.70

quiz(what do you think) What will be the training time complexity of kNN ?

1 user has participated



A  $O(nd)$

0%

B  $O(n)$

100%

C  $O(n\log n)$

0%



D  $O(1)$

0%

[End Quiz Now](#)



OP  
2  
OM PRAKASH S  
3/3 + 257.50



SA  
1  
Souvik Adhikary  
3/3 + 282.26



GV  
3  
G Vibu Vignesh  
3/3 + 244.59

4	NB	Narendra Babu	3/3 ⚡ 232.2
5	KH	Kiran Hebasur	3/3 ⚡ 231.5
6	AS	Anurag Srivastava	3/3 ⚡ 230.3
7	AS	Ayush sachan	2/3 ⚡ 193.7
8	SJ	SHASHANK JHA	2/3 ⚡ 191.3
9	NP	Nachiket Pawar	2/3 ⚡ 190.2
10	AD	Aditya Shandilya	2/3 ⚡ 188.1

## Testing Time Complexity KNN

1. For each point:
  - i. Calculate distance  $\sim O(n \times d)$
  - ii. Append distance to data frame/list  $\sim$  Constant
2. Sort the distance.  $\sim O(n \log n)$
3. Pick top K nearest neighbours. And voting  $\sim O(1)$

$$O(nd) + O(n \log n)$$

↙

Tim Sort  $\rightarrow$  Quick Sort + Merge Sort  
 $< 15$                        $> 15$

Test time complexity

Step 1: Find distance b/w training data and  $x_q = O(n \times d)$

Step 2: Sort data  $= O(n \log n)$

Step 3: Pick nearest neighbour  $O(k)$

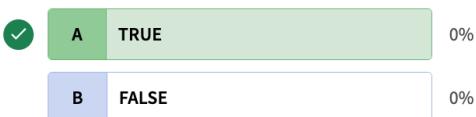
Step 4: Majority vote  $O(k)$



As  $k \ll n$  &  $d$ , hence  $O(k)$  ignored  
Time complexity =  $O(nd + n \log n)$

**k-NN algorithm does more computation on test time rather than train time.**

0 users have participated



[End Quiz Now](#)

 ON PRAKASH S 4/4 346.33	 Souvik Adhikary 4/4 379.40	 G Vibu Vignesh 4/4 333.01
 Narendra Babu 5/5 326.40	 Kiran Hebasur 4/4 317.53	
 Anurag Srivastava 5/5 312.80		
 SHASHANK JHA 5/5 286.59		
 Aditya Shandilya 3/4 284.53		
 Nachiket Pawar 3/4 284.17		
 Periseta Pavan Kalyan 3/4 282.63		

### kNN for Categorical Features

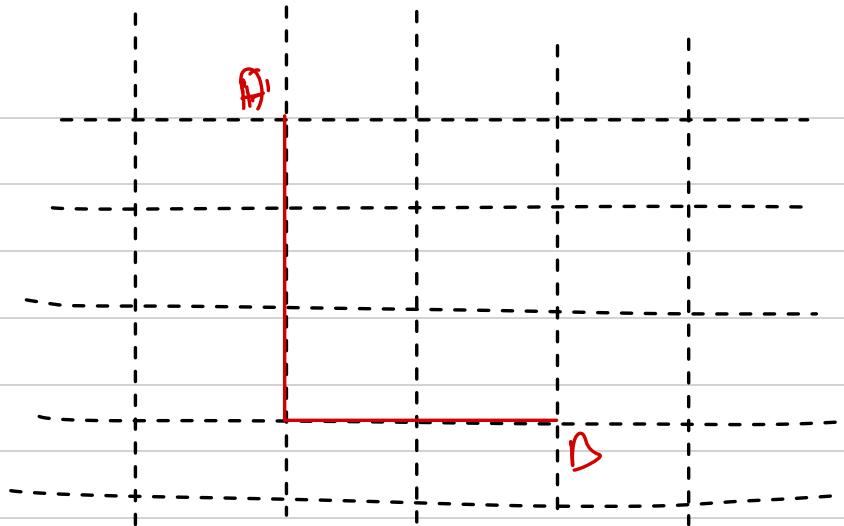
You can't use label-encoding while using KNN for nominal data, you can use however, target-encoding, one-hot encoding, label-encoding (only for ordinal data)

Euclidean distance doesn't work for higher dimensions, we can use:

1. Manhattan
2. Cosine Similarity.

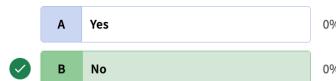
Thumb Rule

$d < 7 \rightarrow$  Euclidean.  
 $15 > d > 7 \rightarrow$  Manhattan  
 $> 15 \rightarrow$  Cosine.



quiz (what do you think) if One Hot Encoding increases data dimension to (d=1000), will Eculidean Distance work ?

0 users have participated



[End Quiz Now](#)

#### Leaderboard

Based on all quizzes from the session

	<b>OM PRAKASH S</b>	5/5	441.92	5/5	406.77
	<b>Souvik Adhikary</b>	5/5	475.10	5/5	404.66
	<b>Narendra Babu</b>	5/5	412.80	4/5	379.06
4	G Vibu Vignesh	5/5	378.00	4/5	378.00
5	Kiran Hebansur	5/5	372.16	4/5	372.16
6	Aditya Shandilya	4/5	370.56	4/5	370.56
7	Nachiket Pawar	4/5	369.36	4/5	369.36
8	SHASHANK JHA	4/5	369.36	4/5	369.36
9	Persettha Pavan Kalyan	4/5	369.36	4/5	369.36
10	Umar	4/5	369.36	4/5	369.36

# Distance metrics

Euclidean distance

$$\left[ \sum_{i=1}^n (m_i - m_j)^2 \right]^{1/2}$$

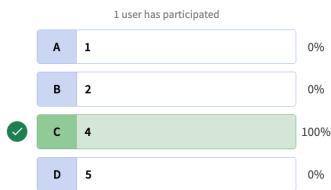
Manhattan

$$\left[ \sum_{i=1}^n |m_i - m_j| \right]^1$$

Minkowski distance

$$\left[ \sum_{i=1}^n |m_i - m_j|^p \right]^{1/p}$$

Which of the following will be Manhattan Distance between the two data point A(1,1,3) and B(1,3,5)?



Leaderboard  
Based on all quizzes from the session

Rank	User	Score
1	OM PRAKASH S	6/6 ₹ 526.62
2	Souvik Adhikary	6/6 ₹ 562.26
3	G Vibu Vignesh	6/6 ₹ 497.41
4	Kiran Hebasur	5/6 ₹ 484.80
5	Aditya Shandilya	5/6 ₹ 472.29
6	Persetta Pavan Kalyan	5/6 ₹ 462.06
7	SHASHANK JHA	5/6 ₹ 461.65
8	Umar	5/6 ₹ 452.23
9	SHISHIR BHAT	5/6 ₹ 443.10
10	Sri Harsha Nanduri	5/6 ₹ 430.76

Clustering - KNNs

## Locality Sensitive Hashing

$w_t$



$r_1$

$r_2$

$r_3$

$r_4$

1

2

3

4

5

6

7

8

9

10

11

12

13

$14+$

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

11

1

2

3

4

5

6

7

8

9

10

## KNN Imputation

$$l^2 = \underline{2}$$

	$F_1$	$F_2$	$F_3$	$F_4$	$F_5$	$F_6$	$y$
$n_1$							
$n_2$							
$n_3$							
$n_4$							
$n_5$							
$n_6$							
$n_7$							
$n_8$							

$$\frac{\text{L2 norm}}{n_1, n_2}$$

$$\text{Red circle} = \frac{\text{Yellow circle} + \text{Green circle}}{2}$$

4, 5

Avatar =

	$F_1$	$F_2$	$F_3$	$F_4$	$F_5$	$F_6$	$y$
$n_1$							
$n_2$							
$n_3$							
$n_4$							
$n_5$							
$n_6$							
$n_7$							
$n_8$							

$$k=3$$

$$S-1$$

$n_2, n_5, n_6$

Made an assumption, that  $x_2, x_5$  and  $x_6$  are top 3 most nearest neighbors.

### Select the true statements

- s1- kNN is less time intensive when LSH is used
- s2- k must be odd
- s3- kNN used for imputing
- s4- For high dimension, euclidean not used

0 users have participated

A	s1	0%
B	s2	0%
C	s3	0%
<input checked="" type="checkbox"/>	D all of the above	0%

[End Quiz Now](#)



OM PRAKASH S  
7/7 605.29



Souvik Adhikary  
7/7 652.69



Kiran Hebasur  
7/7 556.73

4	SHASHANK JHA	6/7 549.95
5	Aditya Shandilya	6/7 546.99
6	Umar	6/7 539.66
7	Sri Harsha Nanduri	6/7 510.66
8	bala chandar kumar	6/7 499.88
9	G Vibu Vignesh	6/7 497.41
10	Anurag Srivastava	6/7 478.43

0.99

$F_1$ ,  $F_2$ ,  $F_3$

$$R_1 \quad \begin{pmatrix} 1 \\ -2 \end{pmatrix}^L \quad R_2 \quad \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix}^L \quad 2 \\ 3$$

$$(R_1 - R_2)^2 =$$