

Session - 2

KNN-2



AGENDA

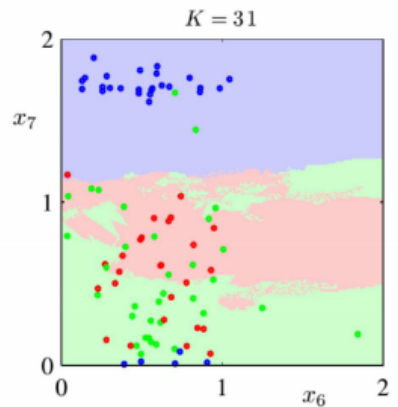
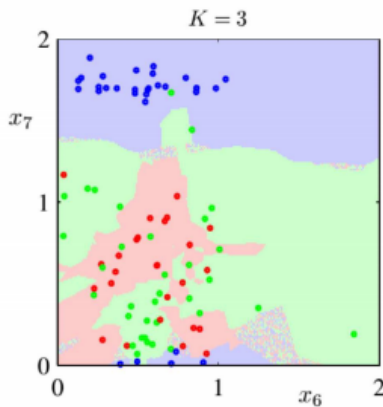
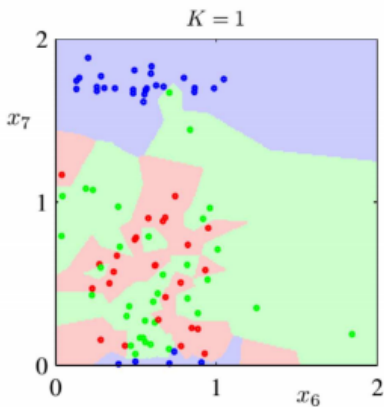
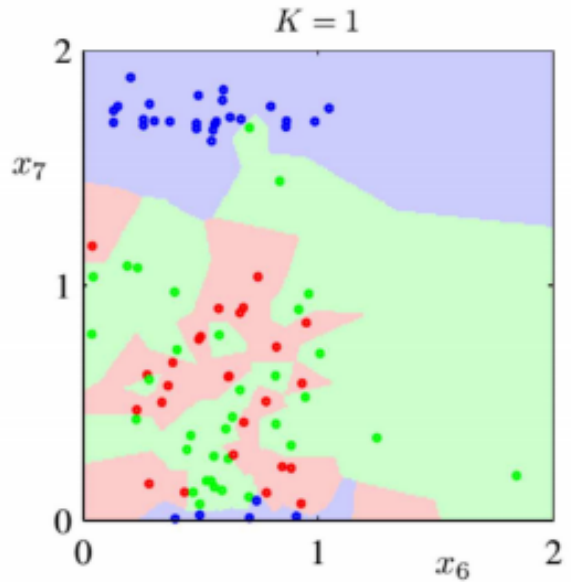
- ① Recap
- ② BIAS- VARIANCE
- ③ HYPER PARAMETER TUNING
- ④ KNN FOR CATEGORICAL FEATURES
- ⑤ DISTANCE METRICS
- ⑥ KNN IMPUTATION

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.



$K = 100$

101
/

Gr - 40
Red - 60

Sort

$P_1 - d_1$

$P_6 \rightarrow d_6$

$P_2 - d_2$

$P_7 \rightarrow d_7$

\rightarrow

,

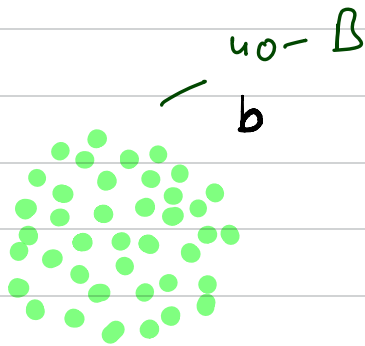
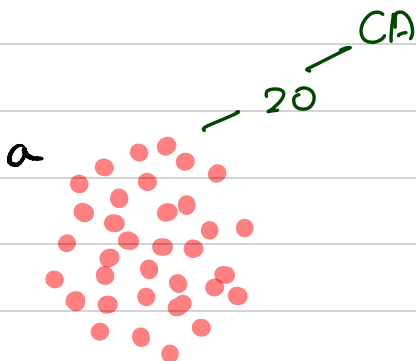
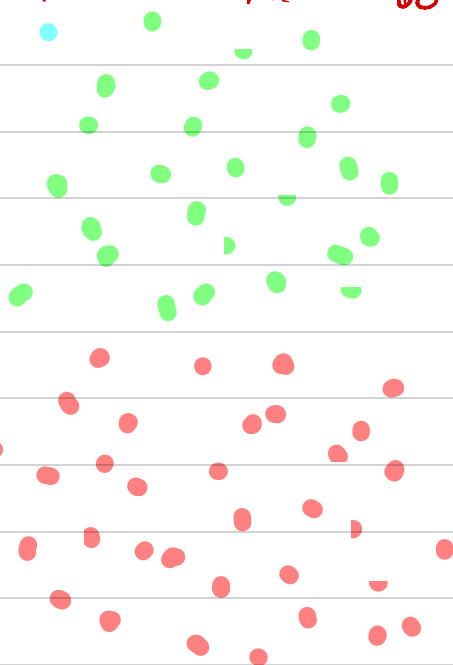
,

,

$P_{100} - d_{100}$

P_1

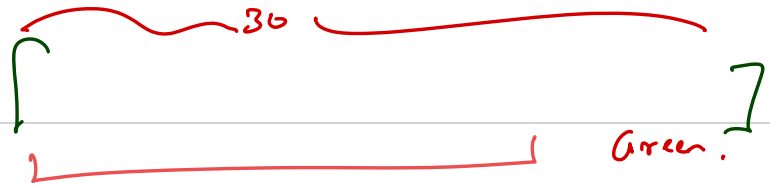
d_1



n_r

$K = \underline{30}$

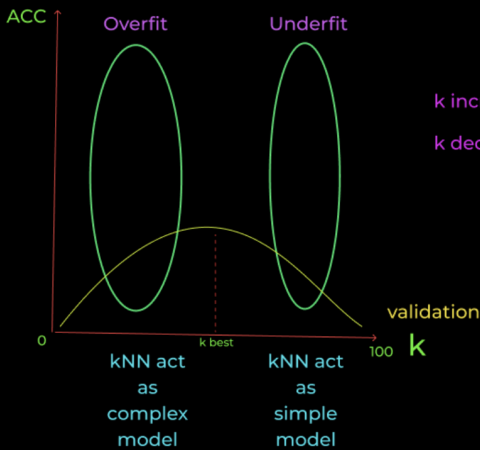
Sorted
dist



$K = 50$



Summary!



k increases, kNN underfits

k decreases, kNN overfits



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

0 users have participated

✓

A

Model overfits

0%

B

Model underfits

0%

End Quiz Now

PC

2

Pankaj Chaudhary

1/1 ⚡ 97.10

AS

3

Aayush sachan

1/1 ⚡ 97.00

AS

1

Aditya Shandilya

1/1 ⚡ 97.36

4	Shoreya gupta	1/1 ⚡ 96.90
5	Vishwajit Verma	1/1 ⚡ 96.87
6	SA Souvik Adhikary	1/1 ⚡ 96.17
7	Nachiket Pawar	1/1 ⚡ 95.93
8	SJ SHASHANK JHA	1/1 ⚡ 95.63
9	PP Perisetla Pavan Kalyan	1/1 ⚡ 95.33
10	S SHISHIR BHAT	1/1 ⚡ 94.70

k→ hyperparameter, then what data to use for hyperparamter 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

Aayush sachan

2/2 ⚡ 193.70

AS

3

Nachiket Pawar

2/2 ⚡ 190.20

4	Aditya Shandilya	2/2 ⚡ 188.13
5	Souvik Adhikary	2/2 ⚡ 187.53
6	PP Perisetla Pavan Kalyan	2/2 ⚡ 186.86
7	Karthik	2/2 ⚡ 182.26
8	GV G Vibu Vignesh	2/2 ⚡ 172.55
9	OM PRAKASH S	2/2 ⚡ 166.96
10	Saklur 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(nlogn)

0%

D

O(1)

0%

End Quiz Now

OM

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	Narendra Babu	3/3 ⚡ 238.2
5	Kiran Hebasur	3/3 ⚡ 231.5
6	Anurag Srivastava	3/3 ⚡ 230.3
7	AS Aayush sachan	2/3 ⚡ 193.7
8	SJ SHASHANK JHA	2/3 ⚡ 191.3
9	Nachiket Pawar	2/3 ⚡ 190.2
10	AS Aditya Shandilya	2/3 ⚡ 188.1

Testing Time Complexity KNN

- For each point:
 - Calculate distance — $O(nd)$
 - Append distance to dataframe/list — Constant
- Sort the distance. — $O(n \log n)$
- Pick top K nearest neighbours. And voting — $O(1)$

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

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

0 users have participated

<input checked="" type="radio"/>	A	TRUE	0%
<input type="radio"/>	B	FALSE	0%

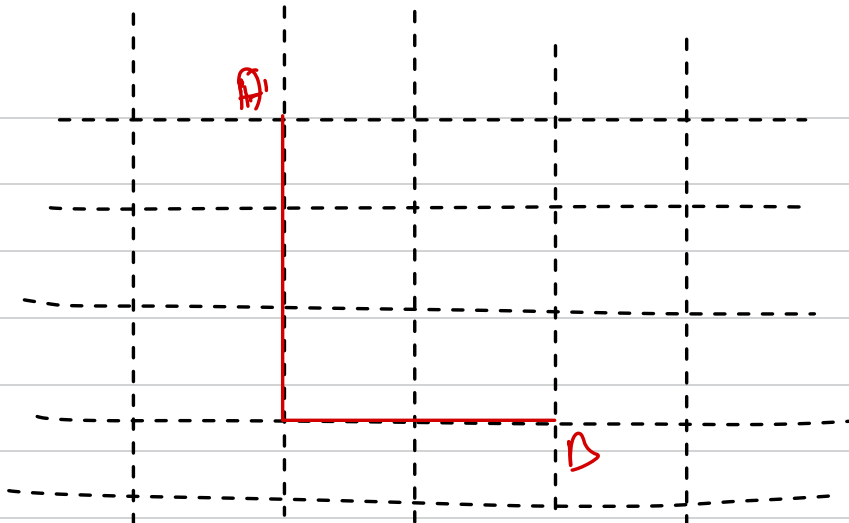
	OM PRAKASH S	Souvik Adhikary	GV Vibhu Vignesh
1	Narendra Babu	4/4	326.40
2	Kiran Hebbar	4/4	317.53
3	Anurag Srivastava	4/4	312.80
4	SHASHANK JHA	3/4	286.59
5	Aditya Shandilya	3/4	284.53
6	Nachiket Pawar	3/4	284.17
7	Perisetla Pavan Kalyan	3/4	282.63

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)

1. Manhattan
2. Cosine Similarity.

Thumb Rule

$$\begin{array}{ll} d < 7 \rightarrow \text{Euclidean.} \\ 15 > d > 7 \rightarrow \text{Manhattan} \\ \geq 15 \rightarrow \text{Cosine.} \end{array}$$



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

0 users have participated

A

Yes

0%

✓

B

No

0%

End Quiz Now

Leaderboard

Based on all questions from this session

SA

1

SA

OM PRAKASH S

5/5

441.92

N

3

N

Souvik Adhikary

5/5

475.10

N

3

N

Narendra Babu

5/5

412.80

4	G Vibhu Vignesh	5/5	406.77
5	Kiran Hebassur	5/5	404.66
6	Aditiya Shandilya	4/5	379.06
7	Nachiket Pawar	4/5	378.00
8	SHASHANK JHA	4/5	372.16
9	Perisetla Pavan Kalyan	4/5	370.56
10	Umar	4/5	369.36

Distance metrics

Euclidean distan

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

Manhattan

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

Min/Max/Ri
distance

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











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

1 user has participated

A	1	0%
B	2	0%
C	4	100%
D	5	0%

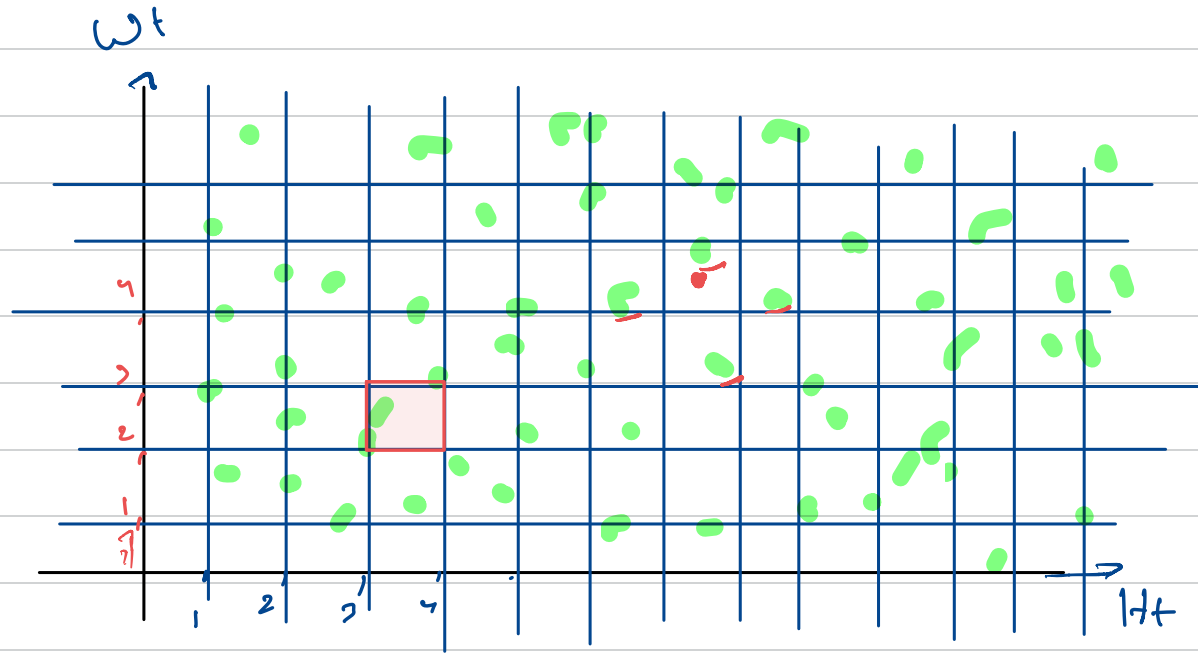
End Quiz Now

Based on all quizzes from the session

 OM PRAKASH S 6/6 ⚡ 526.62	 Souvik Adhikary 6/6 ⚡ 562.26	 GV 6/6 ⚡ 457.41
4  Kiran Hebbsur 6/6 ⚡ 484.80	5  Aditya Shandilya 5/6 ⚡ 472.29	6  Perisetla 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		

Clustered - KNNs

Locality Sensitive Hashing



KNN Imputation

$$k=2$$

$$L_2 \text{ norm}$$

$$\sqrt{x_1^2 + x_2^2}$$

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

$$4, 5$$

Assume =

	f_1	f_2	f_3	f_4	f_5	f_6	y
x_1							
x_2							
x_3							
x_4							
x_5							
x_6							
x_7							
x_8							

$f_1 f_2 f_3 f_4 f_5 f_6 y$

	f_1	f_2	f_3	f_4	f_5	f_6	y
x_1							
x_2							
x_3							
x_4							
x_5							
x_6							
x_7							
x_8							

$$k=3$$

$$S-1$$

$$\{x_2, x_5, x_6\}$$

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

None

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%
- ☒ 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 Vibhu Vignesh	6/7 ⚡ 497.41
10	Anurag Srivastava	6/7 ⚡ 478.43

0.99

F_1, F_2, F_3

$R_1 \begin{pmatrix} 1 \\ - \\ 2 \end{pmatrix}^L \begin{pmatrix} 1.1 \\ - \\ 2.1 \end{pmatrix}^L \begin{matrix} 2 \\ 3 \end{matrix}$

$(R_1 - R_2)^2 =$