Session - 2

KNN - 2

Agenda

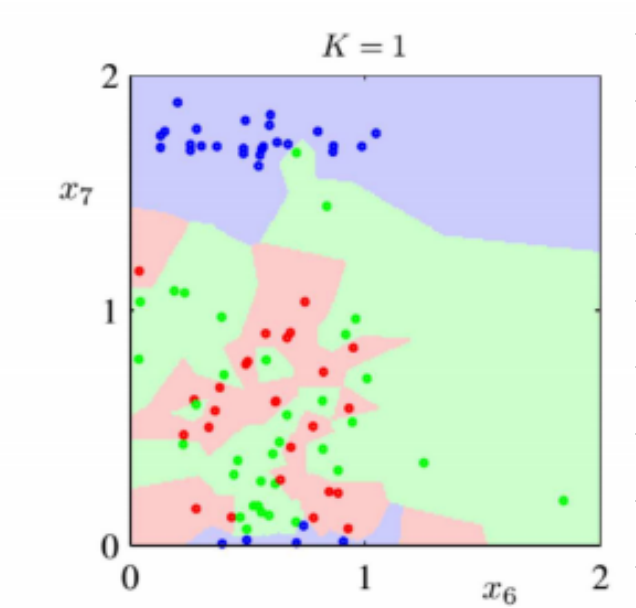
1. Recap
2. Bias - Variance
3. Hyper Parameter Tuning
4. KNN For Categorical Features
5. Distance Metrics
6. KNN Imputation

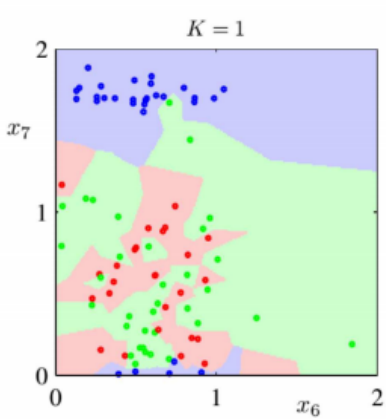
Bias and Variance

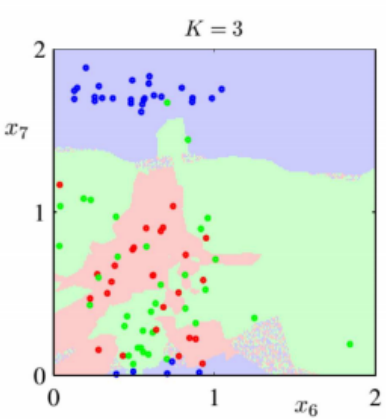
https://datascience.stackexchange.com/questions/81866/why-does-the-overfitting-decreases-if-we-choose-k-to-be-large-in-k-nearest-neigh

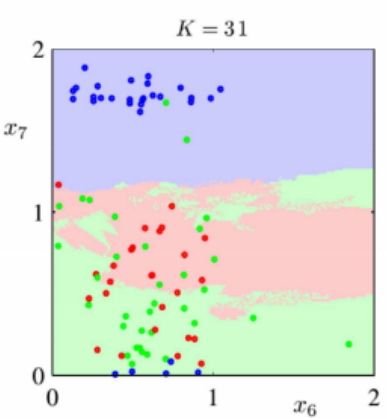
Bias: Errors made by my model on test/production Data because of making very simple assumptions on my training data.

Variance: Errors made by my model on test/ production data because of making very complicated assumptions on my training data.









$K = 100$

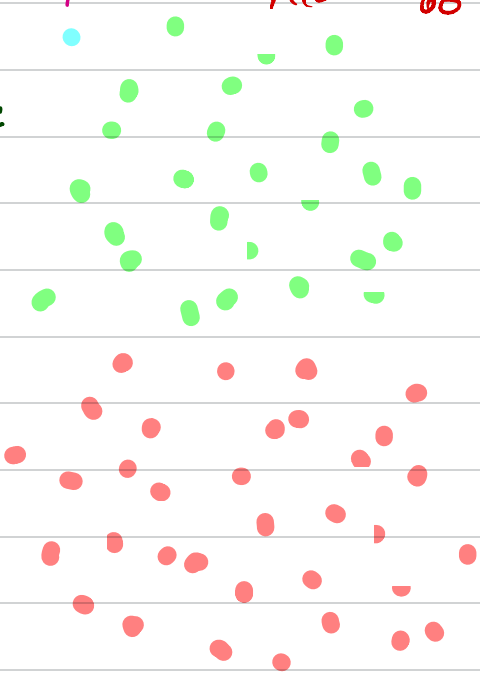
$P_1 - d_1$
$P_2 - d_2$
$\vdots$
$P_{100} - d_{100}$

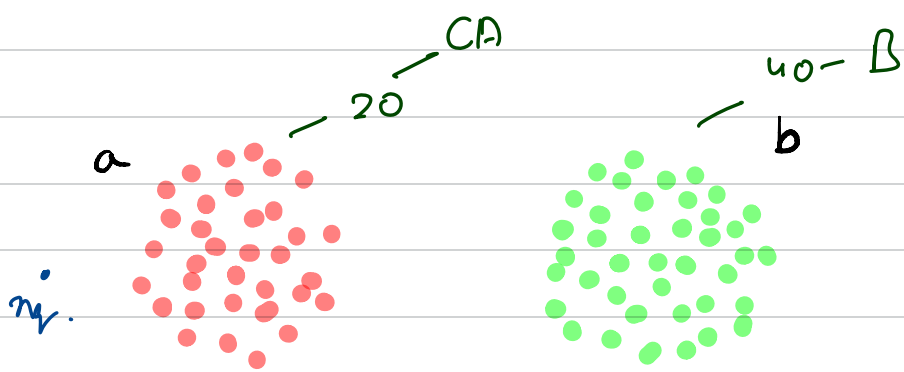
Sort →

$P_6 \to d_6$
$P_7 \to d_7$
$\vdots$
$P_1 \quad d_1$

101

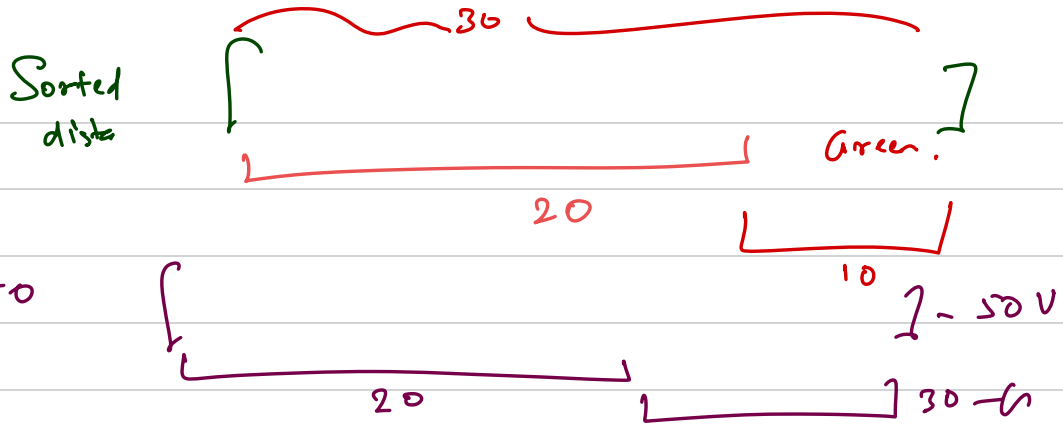
Gr - 40
Red - 60



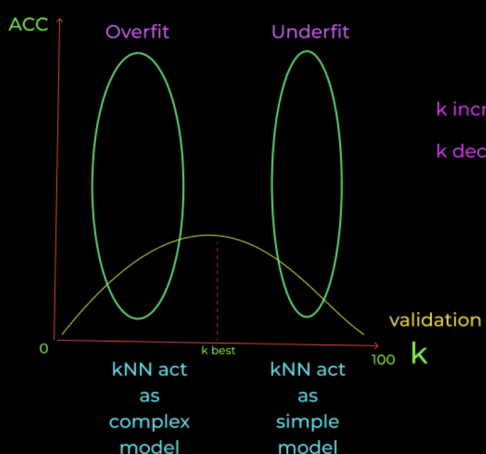


n_y.

$K = 30$



Summary!



k increases, kNN underfits

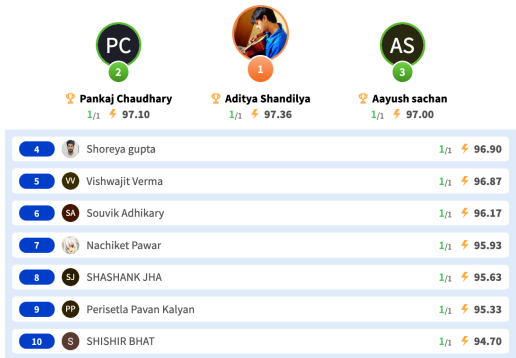
k decreases, kNN overfits

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

0 users have participated

✓ A	Model overfits	0%
B	Model underfits	0%

End Quiz Now

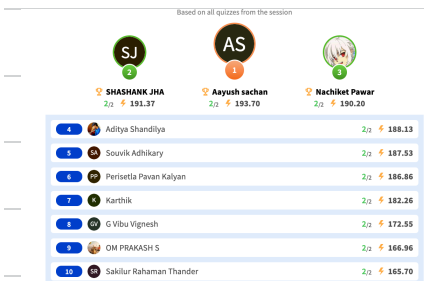


k→ hyperparameter, then what data to use for hyperparamter tuning ?

4 users have participated

✓ A	validation	50%
B	training	50%
C	test	0%
D	entire data	0%

End Quiz Now

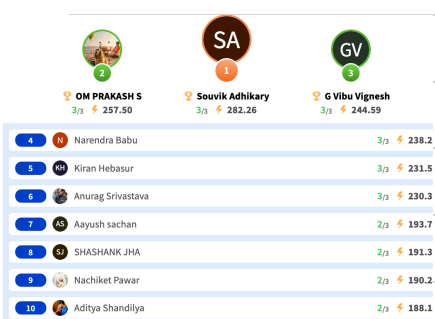


quiz(what do you think) What will be the training time complexity of kNN ?

1 user has participated

A	O(nd)	0%
B	O(n)	100%
C	O(nlogn)	0%
✓ D	O(1)	0%

End Quiz Now



Testing Time Complexity KNN

1. For each point:
   i. Calculate distance — O(n×d)
   ii. Append distance to dataframe/list — Constant
2. Sort the distance. — O(n log n)
3. Pick top K nearest neighbours. And voting — O(k)

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

Tim Sort → Quick Sort + Merge Sort

< 15 > 15

Test time complexity

Step 1 : Find distance b/w training data and xq = O (n x d)

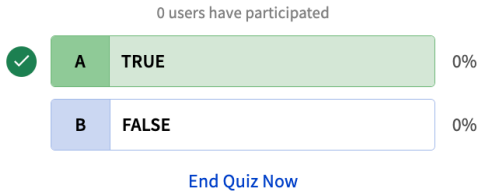
Step 2 : Sort data = O (nlogn)

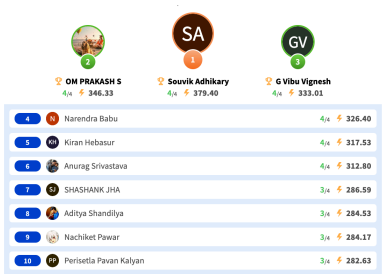
Step 3 : Pick nearest neighbour O(k)

Step 4 : Majority vote O(k)

As k << n & d, hence O(k) ignored
Time complexity = O (nd + nlogn)

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





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:

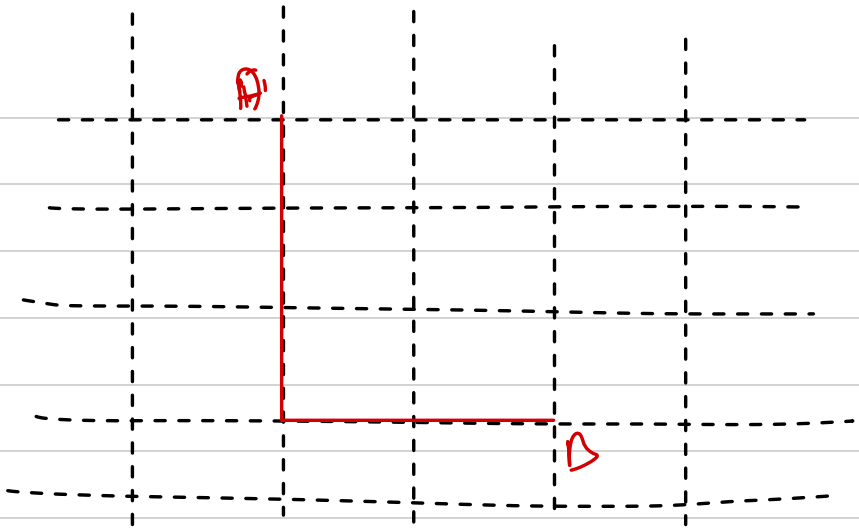
1. Manhattan
2. Cosine Similarity.

Thumb Rule

$d < ?$ → Euclidean.

$15 > d > ?$ → Manhattan

> 15 → Cosine.



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 quizzes from the session

2	1	3
OM PRAKASH S	Souvik Adhikary	Narendra Babu
5/5 ⚡ 441.92	5/5 ⚡ 475.10	5/5 ⚡ 412.80

4	GV	G Vibu Vignesh	5/5 ⚡ 406.77
5	KH	Kiran Hebasur	5/5 ⚡ 404.66
6		Aditya Shandilya	4/5 ⚡ 379.06
7		Nachiket Pawar	4/5 ⚡ 378.00
8	SJ	SHASHANK JHA	4/5 ⚡ 372.16
9	PK	Perisetla Pavan Kalyan	4/5 ⚡ 370.56
10		Umar	4/5 ⚡ 369.36

Distance metrics

Euclidean distance

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

Manhattan

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

Minkowski distance

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

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

1 user has participated

A	1	0%
B	2	0%
✓ C	4	100%
D	5	0%

End Quiz Now

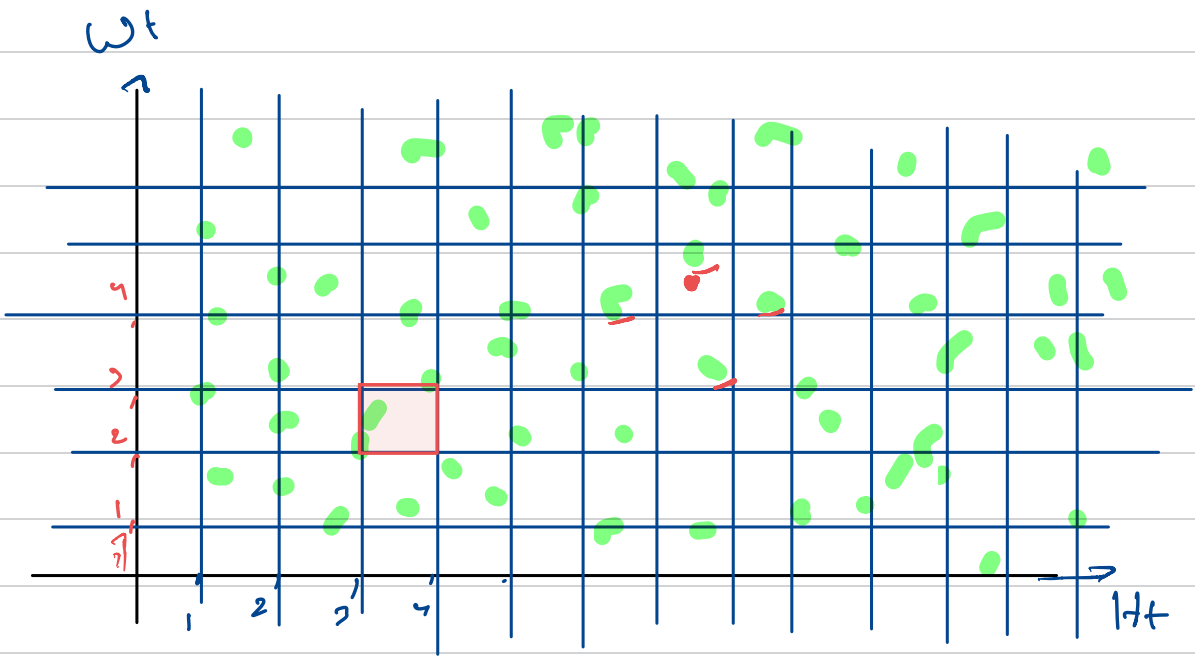
Based on all quizzes from the session

2	SA 1	GV 3
OM PRAKASH S	Souvik Adhikary	G Vibu Vignesh
6/6 ⚡ 526.62	6/6 ⚡ 562.26	6/6 ⚡ 497.41

4	KH	Kiran Hebasur	6/6 ⚡ 484.80
5		Aditya Shandilya	5/6 ⚡ 472.29
6	PP	Perisetla Pavan Kalyan	5/6 ⚡ 462.06
7	S	SHASHANK JHA	5/6 ⚡ 461.65
8		Umar	5/6 ⚡ 452.23
9	S	SHISHIR BHAT	5/6 ⚡ 443.10
10		Sri Harsha Nanduri	5/6 ⚡ 430.76

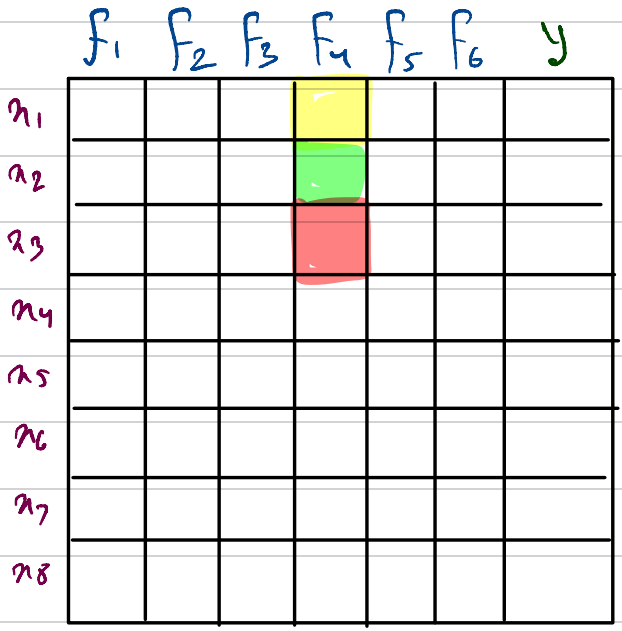
Clustered - KNNs

Locality Sensitive Hashing



KNN Imputation

$k = 2$



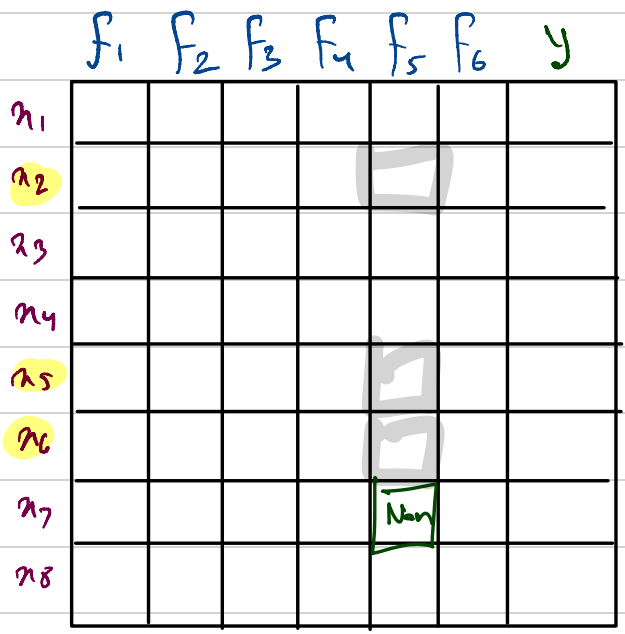
L2 norm

x1, x2

$= \frac{+}{2}$

4 , 5

Avator :



$k = 3$

S-1

x2, x5, x6

Made an assumption, that x2, x5 and x6 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%
✓ D	all of the above	0%

End Quiz Now

0.99

F_1, F_2, F_3

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

R_2 3

$(R_1 - R_2)^2 =$