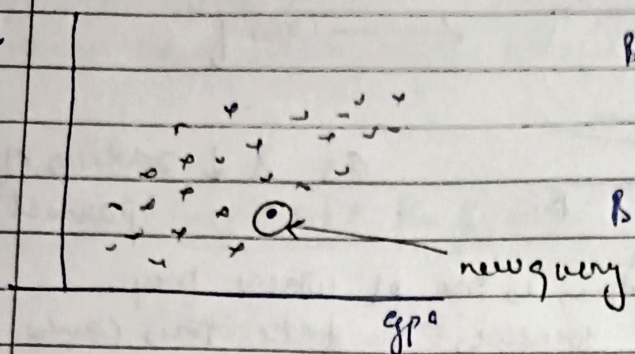
(Black box model) K- Nearest Neighbour (KNN) (classification)



Blue → placement ho gaya (1)

Black → No placement (0)

Suppose a new query comes then we will calculate its distance (euclidean) from all points. Now we sort it in ascending order now do majority count

Lets for pt 1 dist of query → 0 $k=3$
2 → 1
3 → 1

Since among the ~~a~~ 3 nearest pt majority has placed so new query also get placed

• How to select k?

heuristic
→ $k = \sqrt{n}$ n = no. of observation

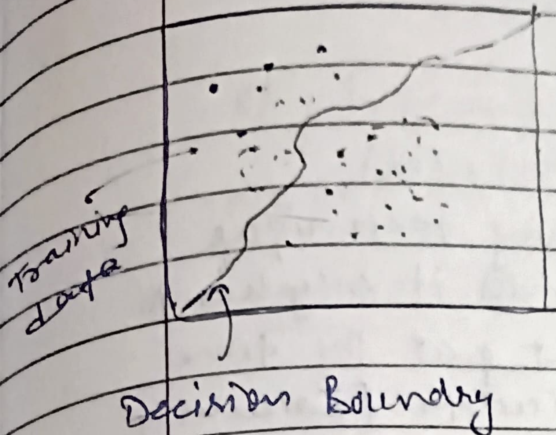
$\sqrt{400} = 20$
↓
don't take even
either take 19 or 21

experimentation
cross validation
$n = 1000$
800 200

apply diff KNN
KNN = 1
2
3
|

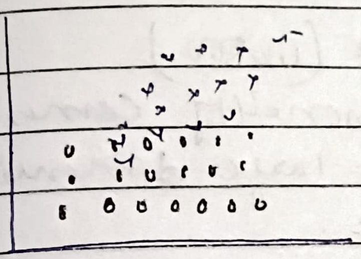
The one which gives best ~~result take~~
that value of K

- Decision Surface
  ↳ It is a tool used in classification algorithm



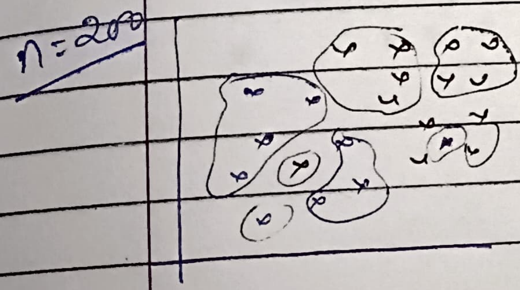
Suppose we considered a 2D data. So There are two surface one for 0 & other for 1.
Now if any new pt. comes, we can tell weather it will fall on region 1 or reg. 0

→ now its made -



→ we plot the training data
→ then generete a numpy meshgrid
→ then apply knn which tells weather 0 or 1
→ Then we make pixels & Show it as an image.

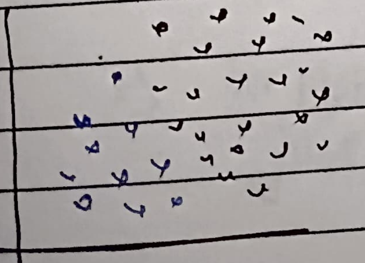
- Overfitting and Underfitting in KNN



K=1

→ This is overfitting ~~as~~
↳ high variance

K=200



↳ If a new pt comes then now we will see its distance from all pts.
Since black is in majority ~~[illegible]~~
So it will always win
→ underfitting

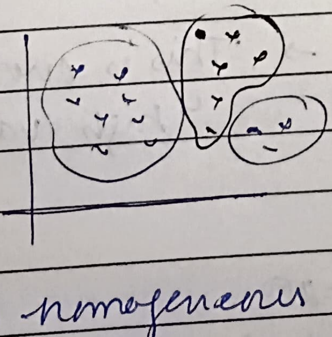
low value of k → overfitting
high " " " → underfitting

- Limitations of KNN-

1) Large dataset .. (500000, 100)
KNN is lazy learning technique means training is fast as it only has to store the points. but ~~&~~ at the time of prediction all calculatn (calculating dist from all pts, sorting etc) happens so it became slow

2) High Dimensional data (like 500)
↳ curse of dimensionality comes into play which says at large dimension dist are not reliable.
Since KNN is based on dist so it is also not reliable

3) Doesn't works good with outliers.



4) Non-homogeneous scales-

exp	sal
0-25	10k - 1lack
small scale	large scale

5) Imbalanced dataset
↳ 98% Yes
↳ 2% No

6) Inference So (Black Box model)
↓
~~not good for~~

· Decision Boundry:-

↳ we can draw decision boundry for all classificatn algo including Neural Network
↳ It can be both Linear or non-Linear
↳ for high dimension problems, it act as an hyperplane.

⇒ Vornoi Diagram- ✓

- Meshgrids

7
y · 6
5
x 1 2 3 4
x

⇒

17 27 37 47
16 26 36 46
15 25 35 45

⇓

1 2 3 4
1 2 3 4
1 2 3 4
xx

7 7 7 7
6 6 6 6
5 5 5 5
yy

⇒ for plotting Decision Boundry just watch lecture