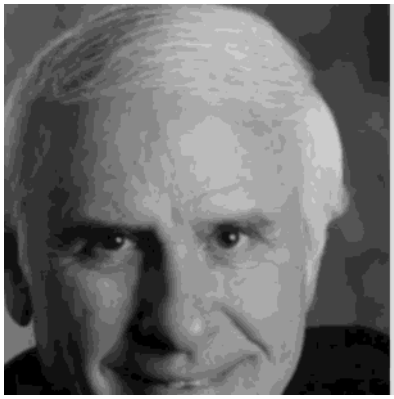
KNN Intuition

22 May 2023 19:14



You are the average of the five people you spend the most time with

- Jim Rohn

KNN → simple elegant

100 student

cgpa	iq	placement
8	80	1
7	70	0
.	.	.
.	.	.

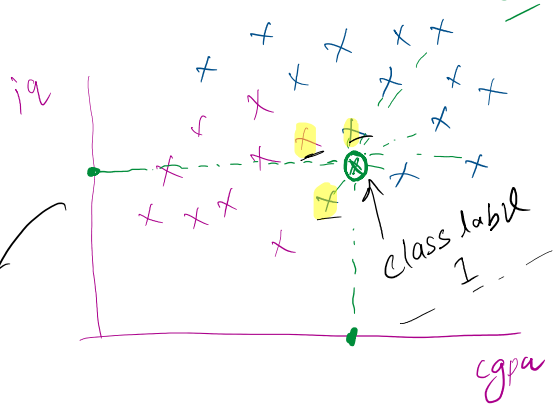
2D

ml model

cgpa → placement
iq

knn

euclidean



2D
↓
n-D

x → 1
x → 0

K = 3

100 distance → 3 nearest neighbors

↳ sort ascending

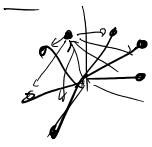
closest

↳ majority count

1 1 0 → 1

$x_1 \mid x_2 \mid x_3 \ldots x_n$

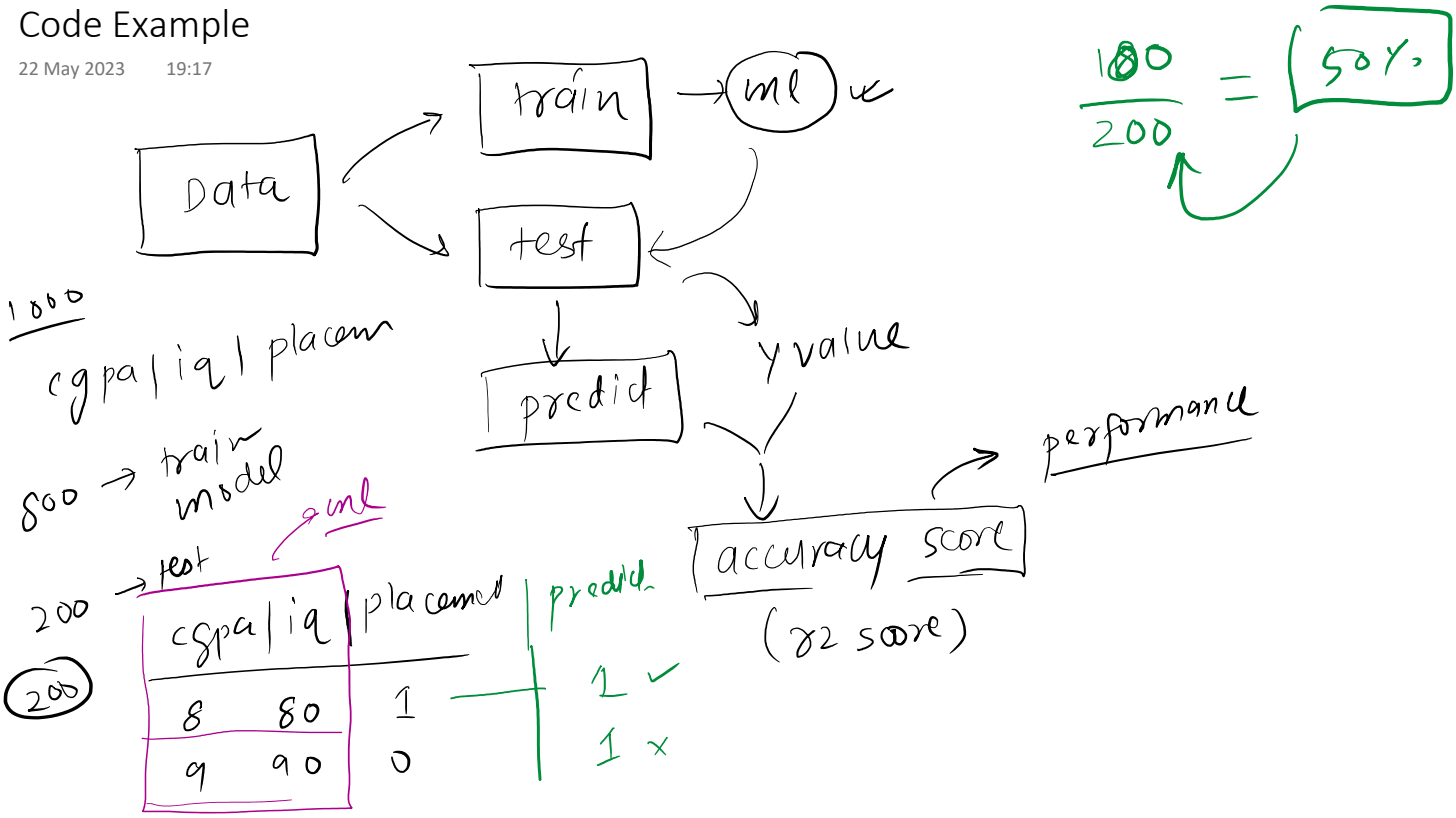
n-dim

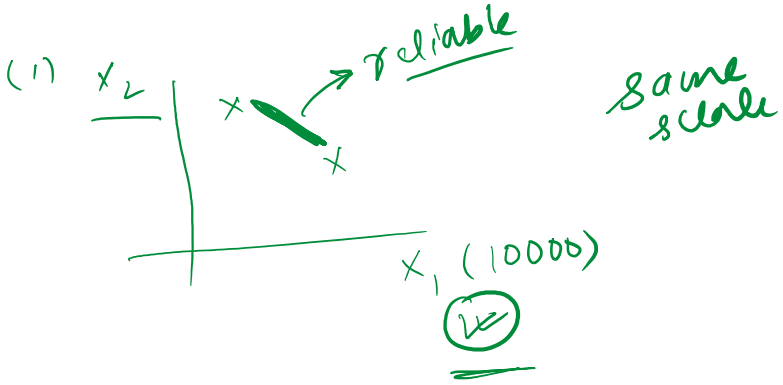


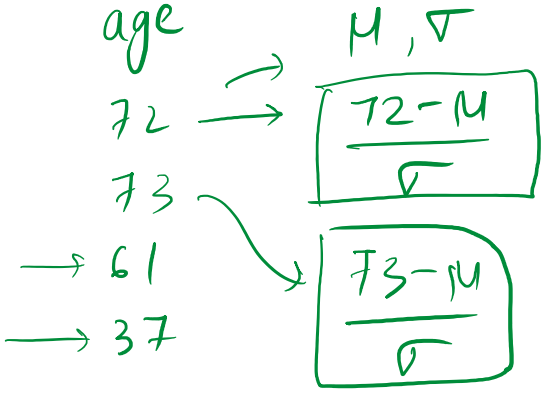
3 vectors → majority count → class label of query point

Code Example

22 May 2023 19:17

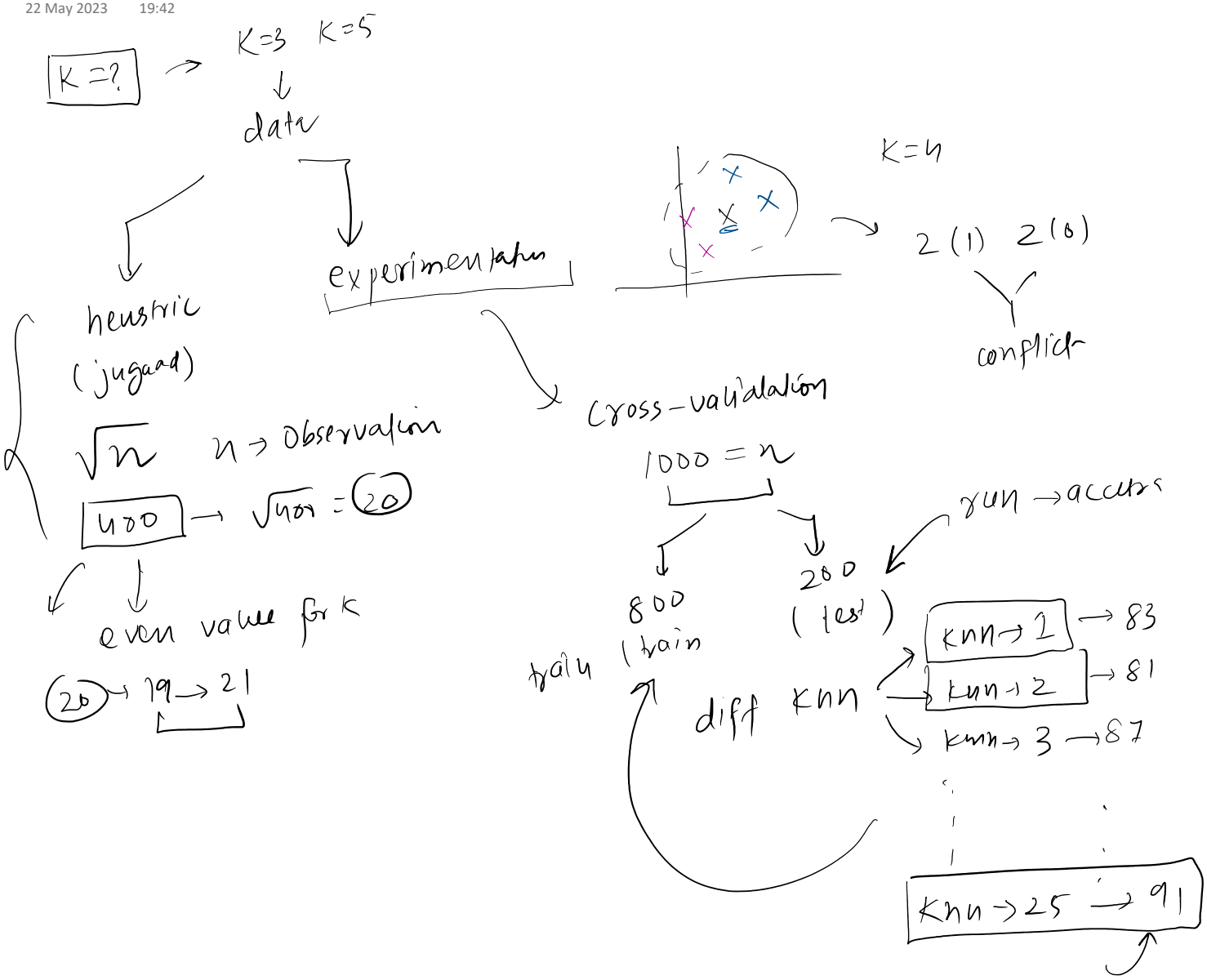






How to select K?

22 May 2023 19:42



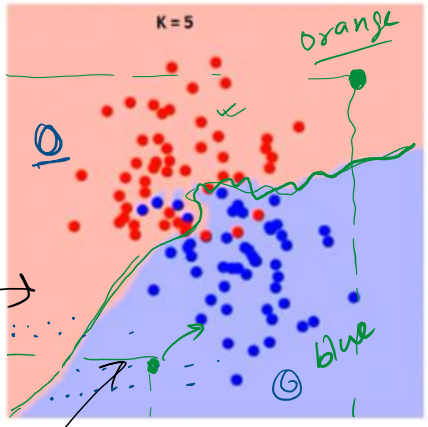
[Decision Surface]

22 May 2023 19:15

→ tool → classification → knn → svm, lr, dt, nn

1D 2D 3D

in



coordinate

disc video

y Dmins → mlxtend → plot decision surfaces

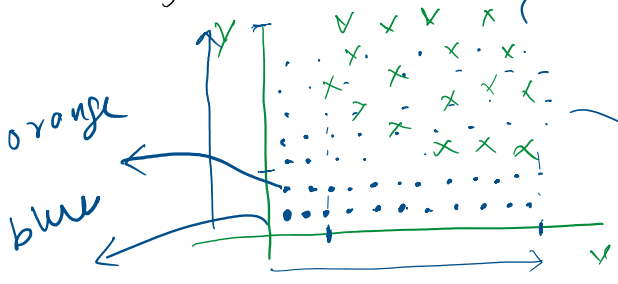
cgpa | iq | place — binary [0,1]

training points

cgpa

pixel on anims

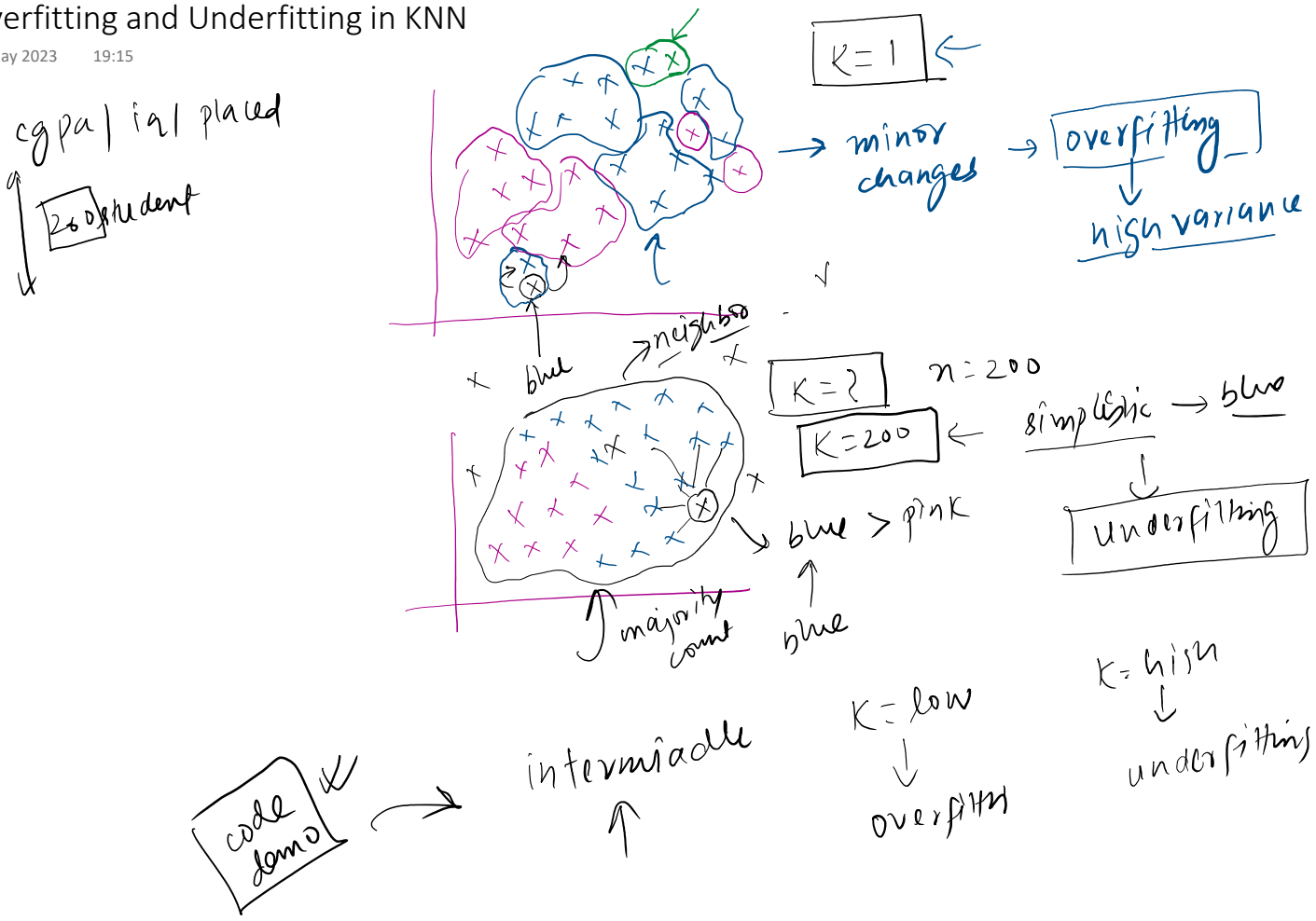
generate a numpy meshgrid



knn train → 1, 0

Overfitting and Underfitting in KNN

22 May 2023 19:15



Limitations of KNN

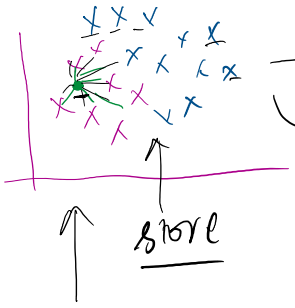
22 May 2023 19:32

failure case

1) large datasets → $n = 5L$, $f = 100$ → (500000, 100)

↳ KNN is a lazy learning tech

prediction → slow → dataset

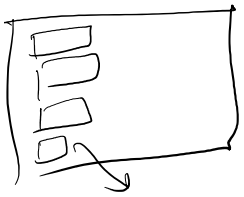


→ query point → prediction

↳ train → nothing

store

5 & distance → sort → majority



low latency

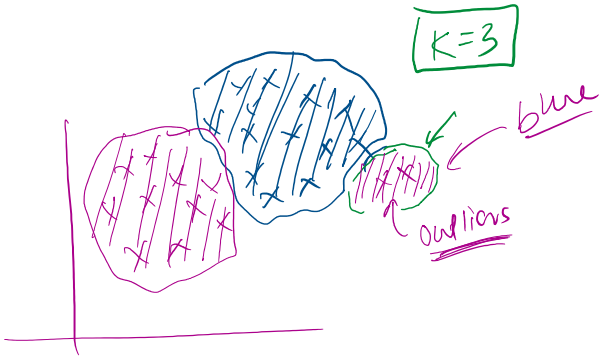
3 sec → slows

2) High dim data → f = 500 →

curse of dimension → distance → reliable concept

KNN

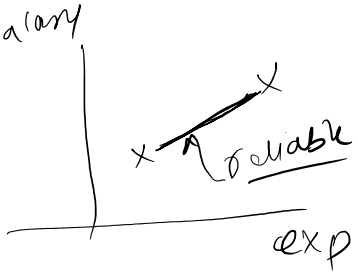
3) Outliers



4) Non-homogeneous scales

dominate

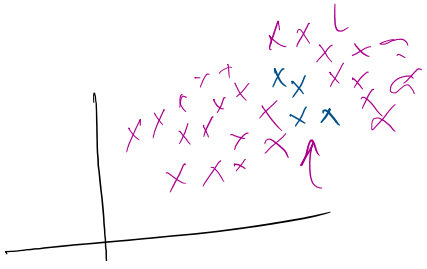
exp	salary	fire
0-25	20K-1Cr	



→ scaling

5) Imbalanced dataset

↳ Yes → 98%

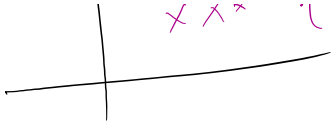


5) Imbalanced

↳ Yes → 98%

No → 2%

↳ biased → print



6) Inference and not for prediction