Lecture : 09

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

K-Nearest Neighbor (KNN)

→ It's supervised learning - classification Algorithm / Regression

Bias: the inability of model to truly capture the relationship in the training data.
(diff b/w our actual and predicted values)

Variance: It measures how much the prediction of a model vary for different training datasets.
(model's sensitivity to fluctuations in the data)

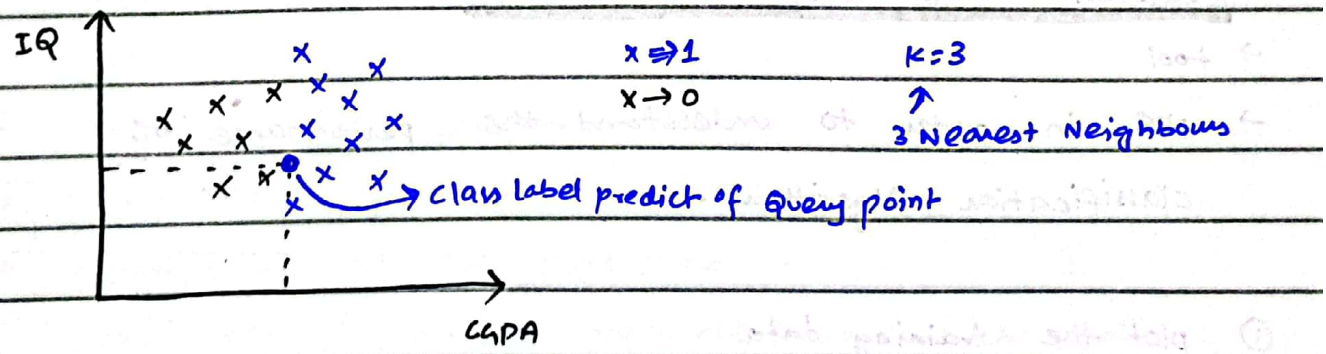
Overfitting: training error low and test error high

Underfitting: training error high

★ (Target would be low bias, low variance)

Example

↝ Let suppose we have CGPA and IQ of students and we have to predict about their placement selection.



→ calculate euclidean distances $d(xy) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$

→ sort distances in ascending order

→ majority count

How to select K ?

① heuristic approach

$\sqrt{n}$ → no of observation

↳ use odd number

(e.g) $n = 400$

$\sqrt{400} = 20$

either 19, 21 for K

② Experimentation approach

(e.g) $n = 1000$

800 training

200 testing

knn → 1

knn → 2

knn → 3

⋮

knn → 25

build different knn models for n = 1, 2, 3 ....

and select the best one

Decision Surface / Decision Boundary :

→ tool

→ use in order to understand the performance of classification Algorithm ..

① plot the training data

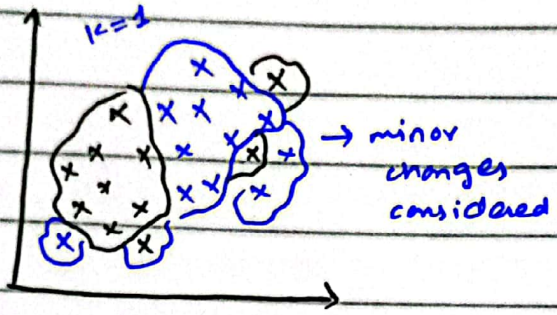
② plot range of training data on x and y-axis

③ for the range generate a numpy meshgrid

Overfitting and underfitting in KNN:

overfitting → high variance (K → low values)

underfitting → K → high values



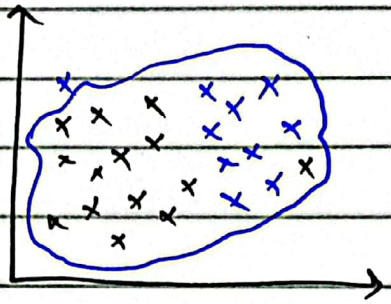
→ minor changes considered

so (overfitting)

(majority count)

→ for a new query point- majority count is considered

so (underfitting)



K = higher

(e.g) K = 200

Limitations of KNN:

→ Large datasets

↳ knn is a lazy learning technique. → Nothing happen in training phase → prediction phase ✓ slow

→ High dimensional data

• if dataset has high dimensions (distance concept is not reliable)

→ outliers

→ Non-homogenous features scale

→ Imbalanced dataset

→ Inference and not for prediction

It does not provide inference (means which feature has high/low impact)

KNN for Regression:

→ Average of neighbours values