17-Nov-2023

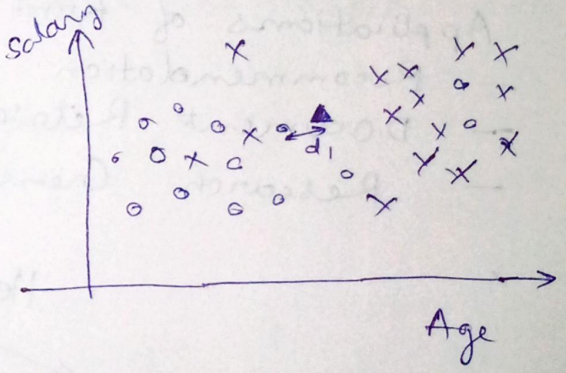
K-Nearest Neighbors

Assumptions in knn:

- knn assumes data is in metric space and there is a notion of distance.
- each of the training data consists of a label data associated with it, either + or -. Although knn also supports multiclass classification.
- We are also given a single number 'k'. This number decides how many neighbors influence the classification. This is usually a odd number.

Geometric intution:

S/N	Age	Salary (thousand)	purchased
1.	25	20	N
2.	63	120	Y
3.	33	75	N
4.	42	100	Y



case-1 $k = 1$ it means neighbors = 1

d_1 = nearest distance from nearest data point

it mean new data point is purchased.

case-2 $k = 3$

here you can see according to graph.
one 'O' (purchase) and two 'X' (not purchased)
hence according to the majority count the new point will be classified as ~~pur~~ not purchased

point to consider

- Although in this example we are talking about a 2D example but the concept holds true for higher dimension as well.
- In this example we have taken euclidean distances into consideration but other distances are used as well like Manhattan distance or Minkowski distance.

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

- The knn is suitable for low noise data.
- The knn is lazy learner. (Discriminative function)
- also Works for Regression.

Applications of knn:

- Recommendation Systems.
- Document Retrieval Systems
- Research Gene expression.

How to find K

method-1

K = sqrt(No. of data in the training set)

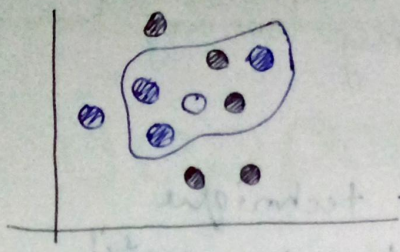
K should be odd to avoid ambiguity

method-2

Trial and Error

Weighted KNN:

Weighted KNN is a variant of KNN where we take a simple yet elegant assumption that the impact of nearer neighbors on the query/test point should be more than the farther away points.



$k = 5$ $\quad wi = 1/di$

point	label	Distance	weight
(x_1, y_1)	Black	0.2	5
(x_2, y_2)	Black.	0.5	2
(x_3, y_3)	Blue	0.7	1.4
(x_4, y_4)	Blue	1.2	0.8
(x_5, y_5)	Blue	1.5	0.6

calculate weight Based on weighing function
Distance Increases, weight decreases simplest weighing function.

- $1.4 + 0.8 + 0.6 = 2.8$
- $5 + 2 = 7$

$7 > 2.8$

Hence label of 'O' is Black.

1. Why KNN is known as a lazy learning technique

lazy learning Technique:
KNN is known as a lazy learning technique because it defers the model's learning until prediction time, making minimal assumptions during training.

Description:
it classifies or predicts based on the majority class or average of the 'k' nearest neighbors in the training data.

Example:
Let's say we have a dataset of flowers with features like petal length and width. When we want to classify a new flower, KNN finds the 'k' training examples with the most similar feature value (nearest neighbors) and assign the majority class among them to the new flower.

Advantages:
- simplicity in implementation.
- Ability to capture complex decision boundaries.
- No need to retain the model when new data arrives.

Disadvantages:
- Computationally expensive for large datasets.
- Sensitive to the choice of 'k'
- prone to noise and outliers

use cases:
- image recognition
- Anomaly detection
- Medical diagnosis
- Recommender systems
- Handwriting recognition