K – Nearest Neighbor Classification

Shania H

Research Scholar

Mehta Family School of Data Science and Artificial Intelligence

IIT Guwahati
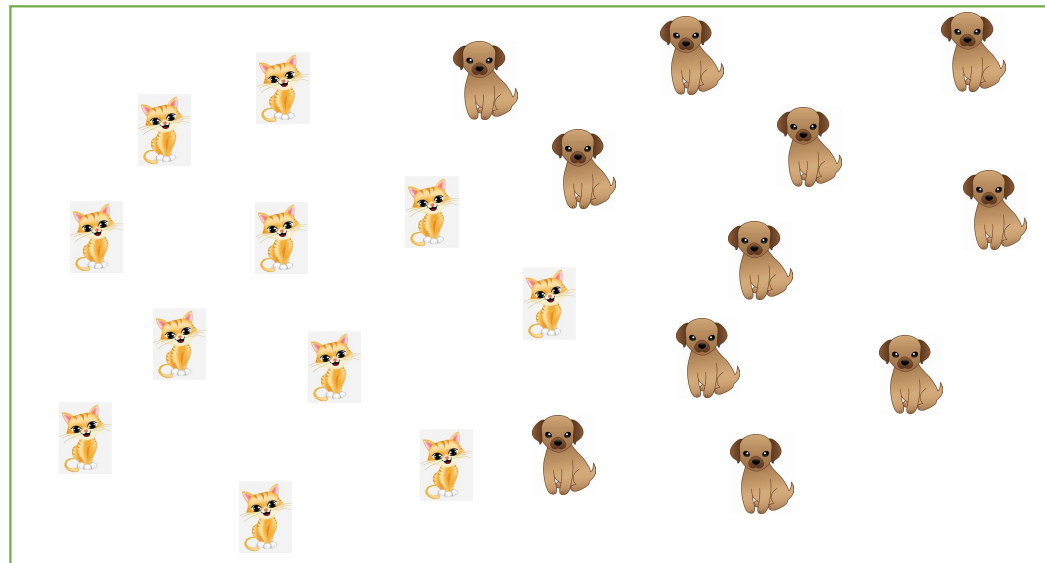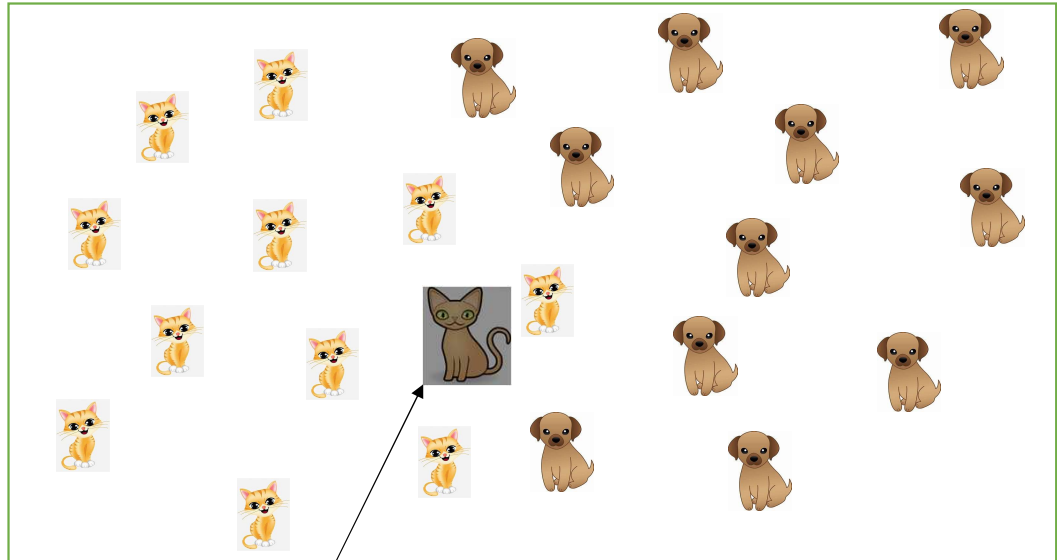
What is KNN?

- Supervised Learning Algorithm
- Lazy – Learner
- Non – linear Learning Algorithm
- Used for both Classification and Regression
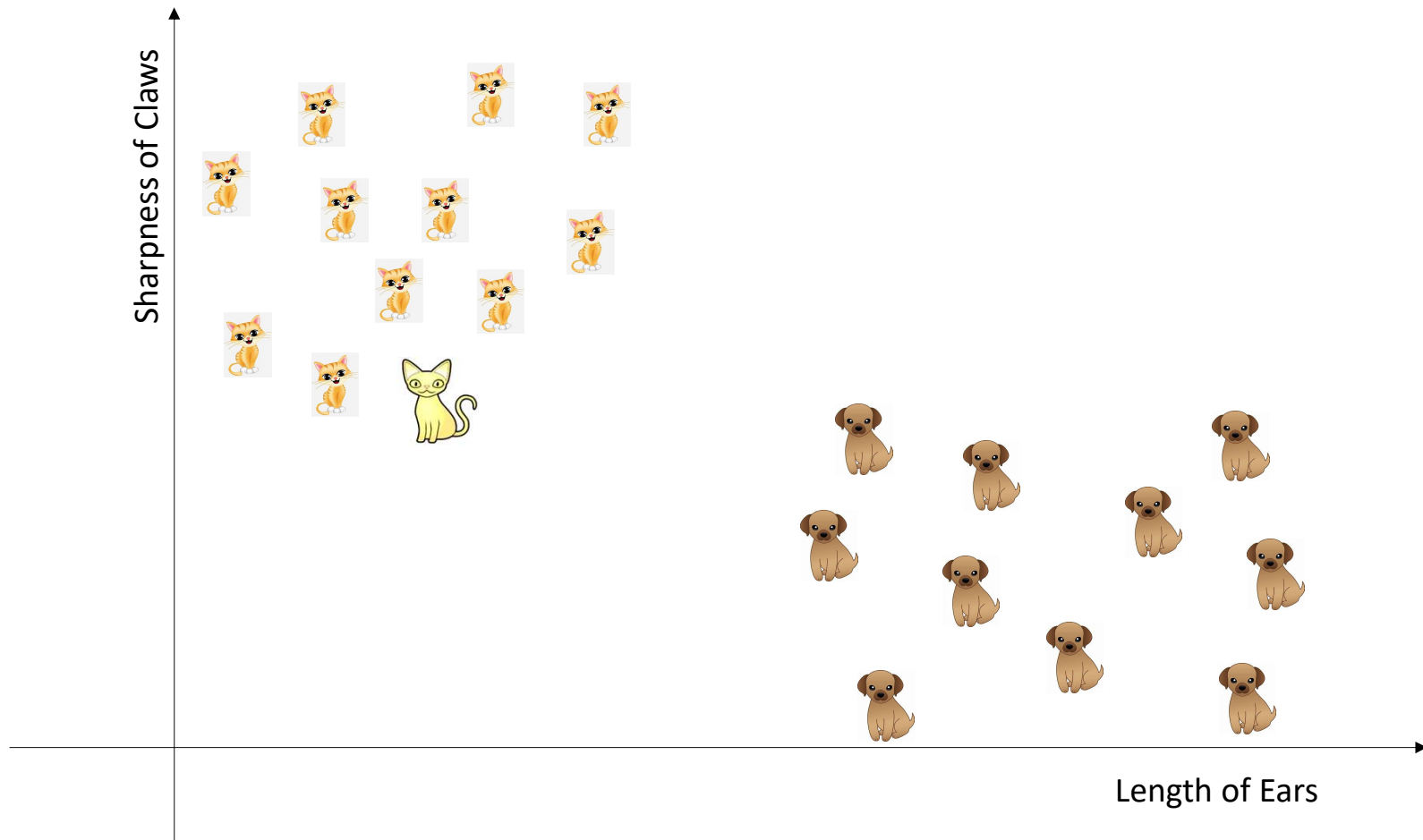
Why KNN Classification?

- Based on feature similarity
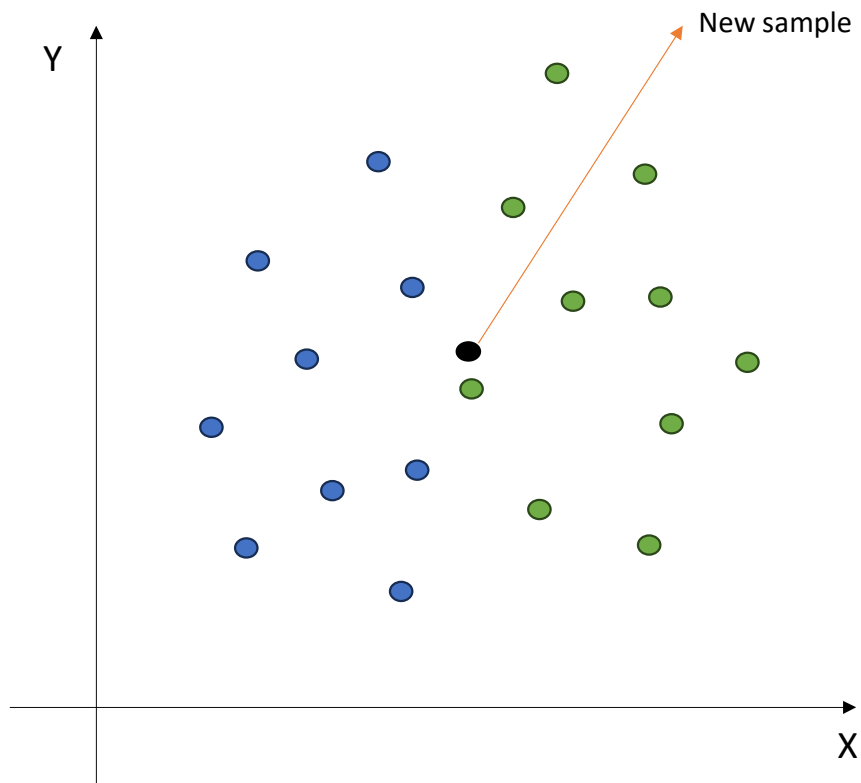
  Example: Classifying Cat and Dog

New Sample

Sharpness of Claws

Length of Ears

- Based on distance metrics

Y

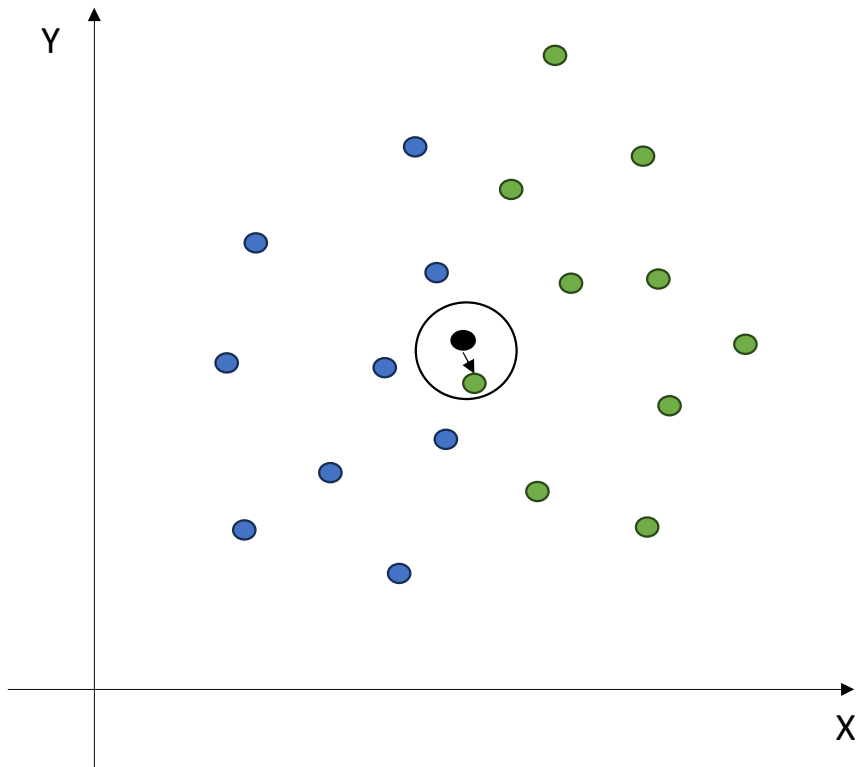New sample

X

Distance Metrics:

✓ Euclidean Distance

✓ Manhattan Distance

✓ Minkowski Distance

Euclidean Distance Formula

For a gives two points (x_1, y_1) and (x_2, y_2), the Euclidean Distance can be calculated as:
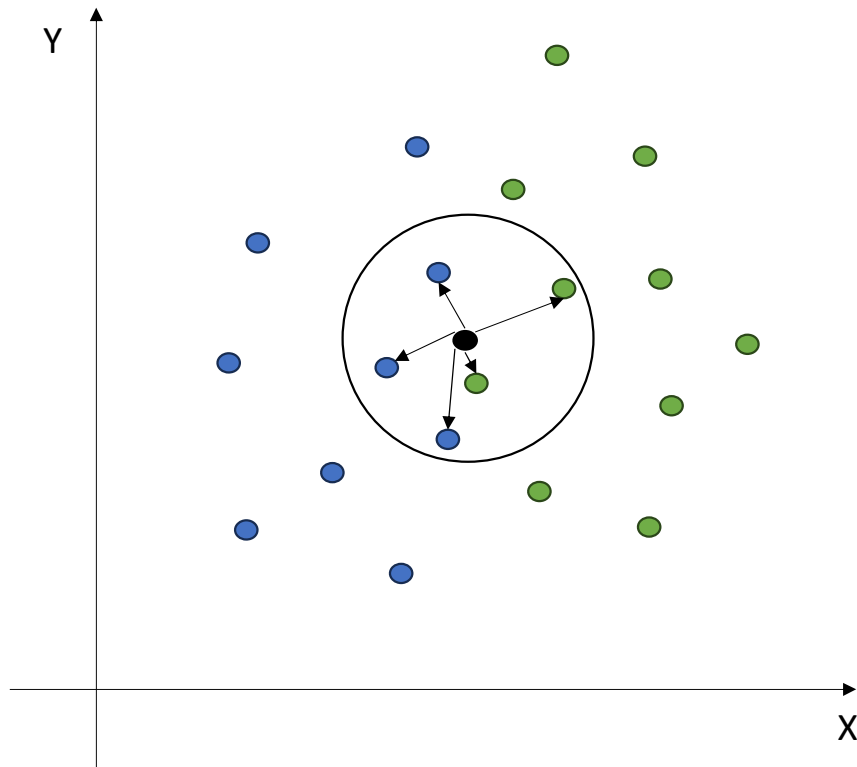
$$Distance, D = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

K = 1

Y

X

When K = 1, there is only one
sample to compare with the new
sample and the new sample has
the shortest distance and it is
classified to Green class.

K = 5

Similarly, when K = , the distance between the new sample and each other samples is found and sorted. Based on the majority of the class, the new sample is classified.
Here, the new sample is classified to Blue class.

How to choose 'K'?

- Choose odd values to avoid ties in classification

- General thumb-rule, $K = \dfrac{\sqrt{N}}{2}$ where N is the number of samples in a training dataset

- If there is a noise in the dataset, higher values of K would be better

- Cross – validation

Advantages

- Easy to understand and implement
- No training required
- Good accuracy
- Can handle large dataset

Disadvantages

- Computationally expensive

- Limited to Euclidean Distance

- Need good choice of 'K'

- Memory requirement for processing on large datasets

- Not suitable for a noisy dataset

References

- *GeeksforGeeks* – KNN Algorithm
- *JavaTpoint* – KNN Algorithm
- *IBM* – KNN Algorithm
- *Medium* – Explanation of KNN Algorithm

THANK YOU!