ML - 2

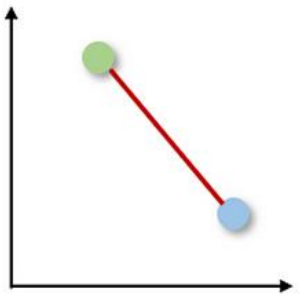
Machine Learning

- Supervised Learning
  - Distance based Classification
    - K-Nearest Neighbors (KNN)

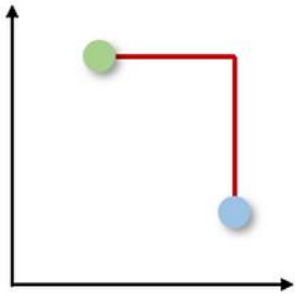
Similarity

Some Similarity Measures

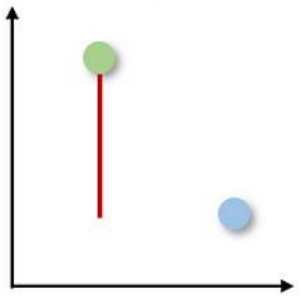
Euclidean



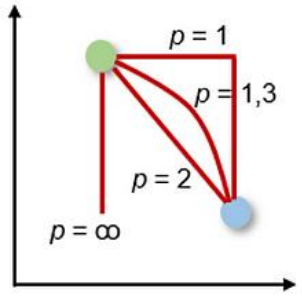
Manhattan



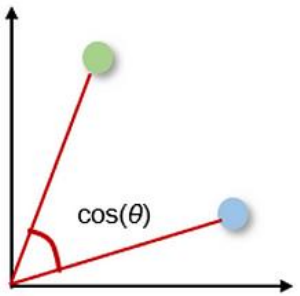
Chebyshev



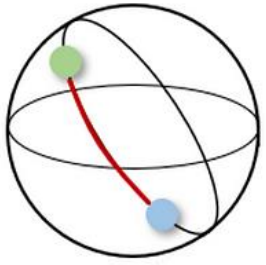
Minkowski



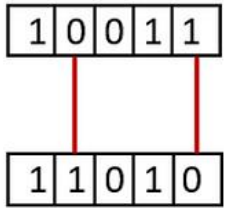
Cosine



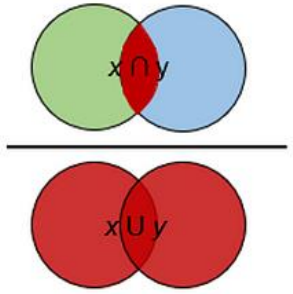
Haversine



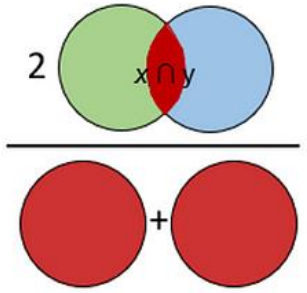
Hamming



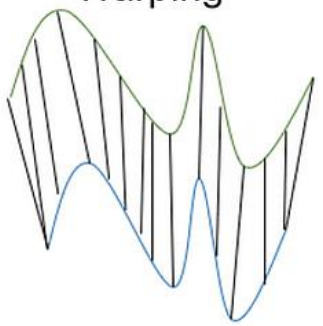
Jaccard



Sörensen-Dice



Dynamic Time Warping



Similarity vs. Distance

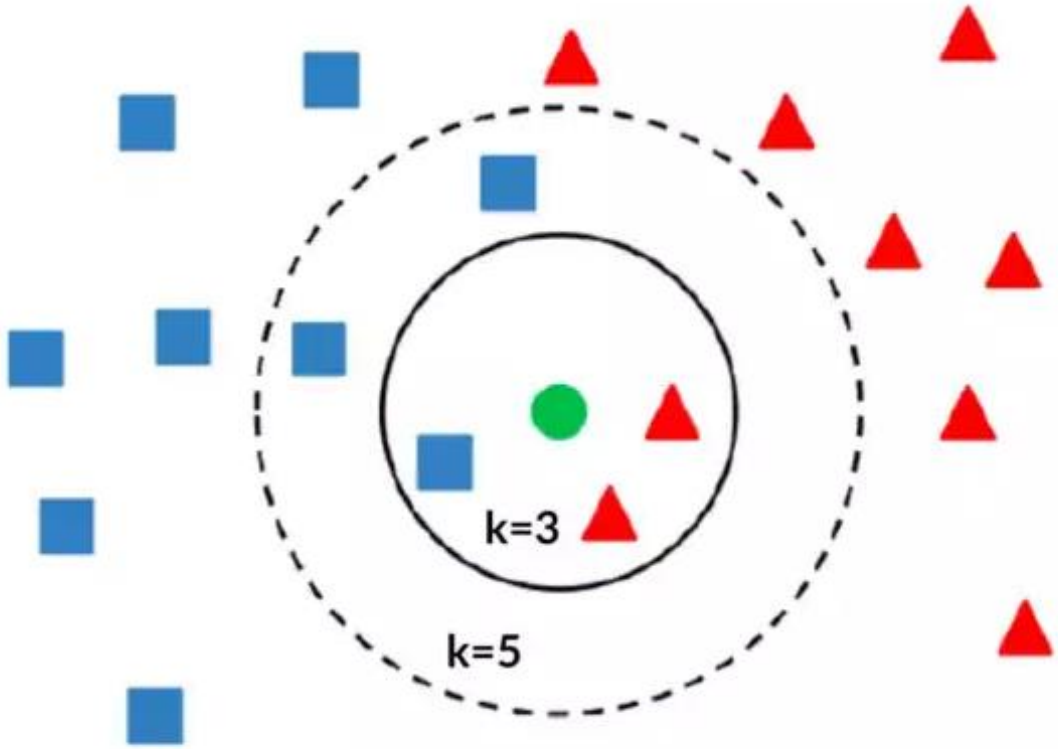
Similarity with inverse of distance:

$$S_{ab} = \frac{1}{1 + \|a - b\|}$$

Similarity with negative exponent of distance:

$$S_{ab} = \frac{1}{e^{\|a-b\|}}$$

K Nearest Neighbor (K-NN)



K-NN Algorithm

- ❑Request K value
- ❑Calculate the distances
- ❑Sort distances
- ❑Select the top K values
- ❑Determine the most frequent class label

A Sample (K-NN)

K=?

A Sample (K-NN)

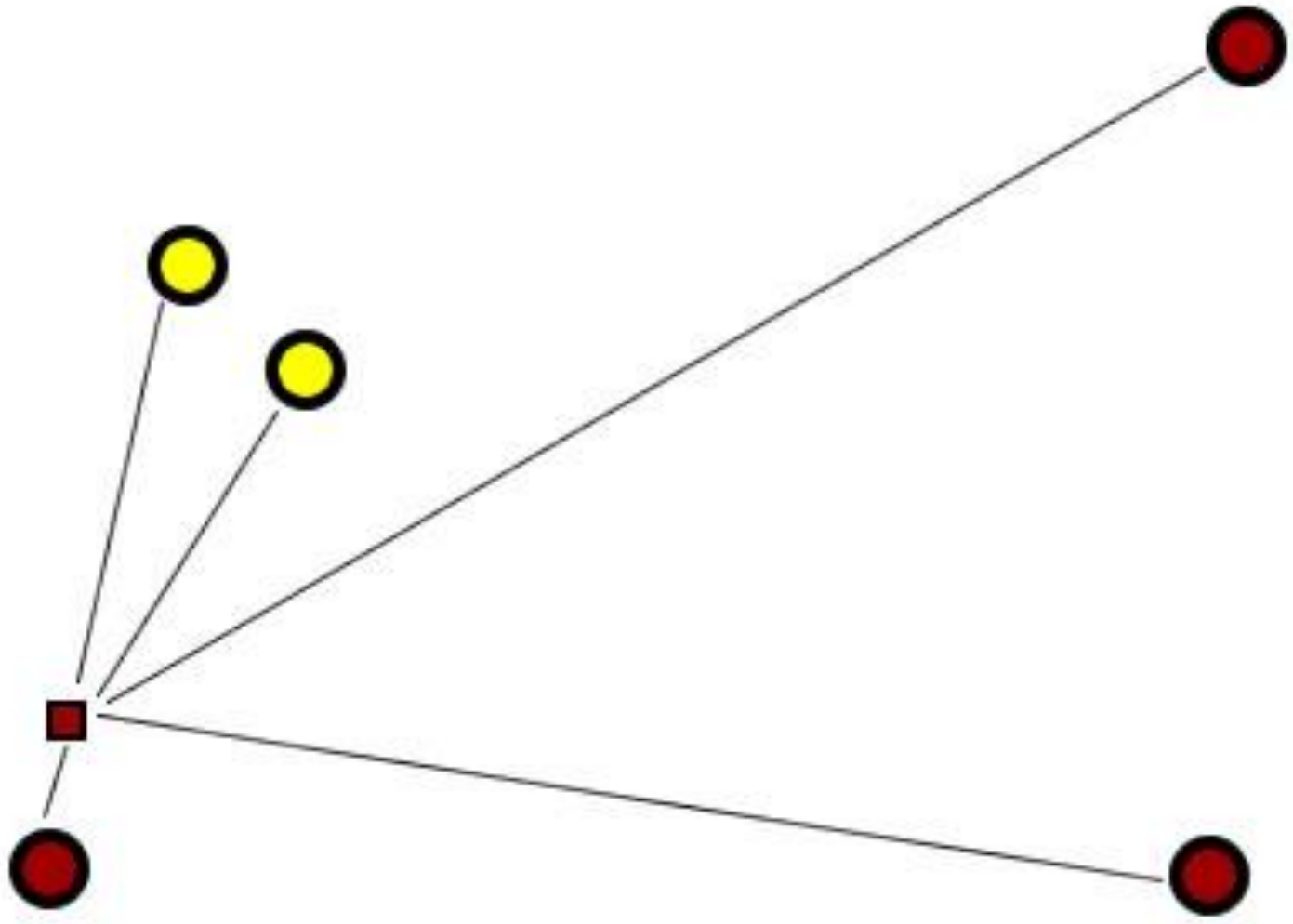
K=1

A Sample (K-NN)

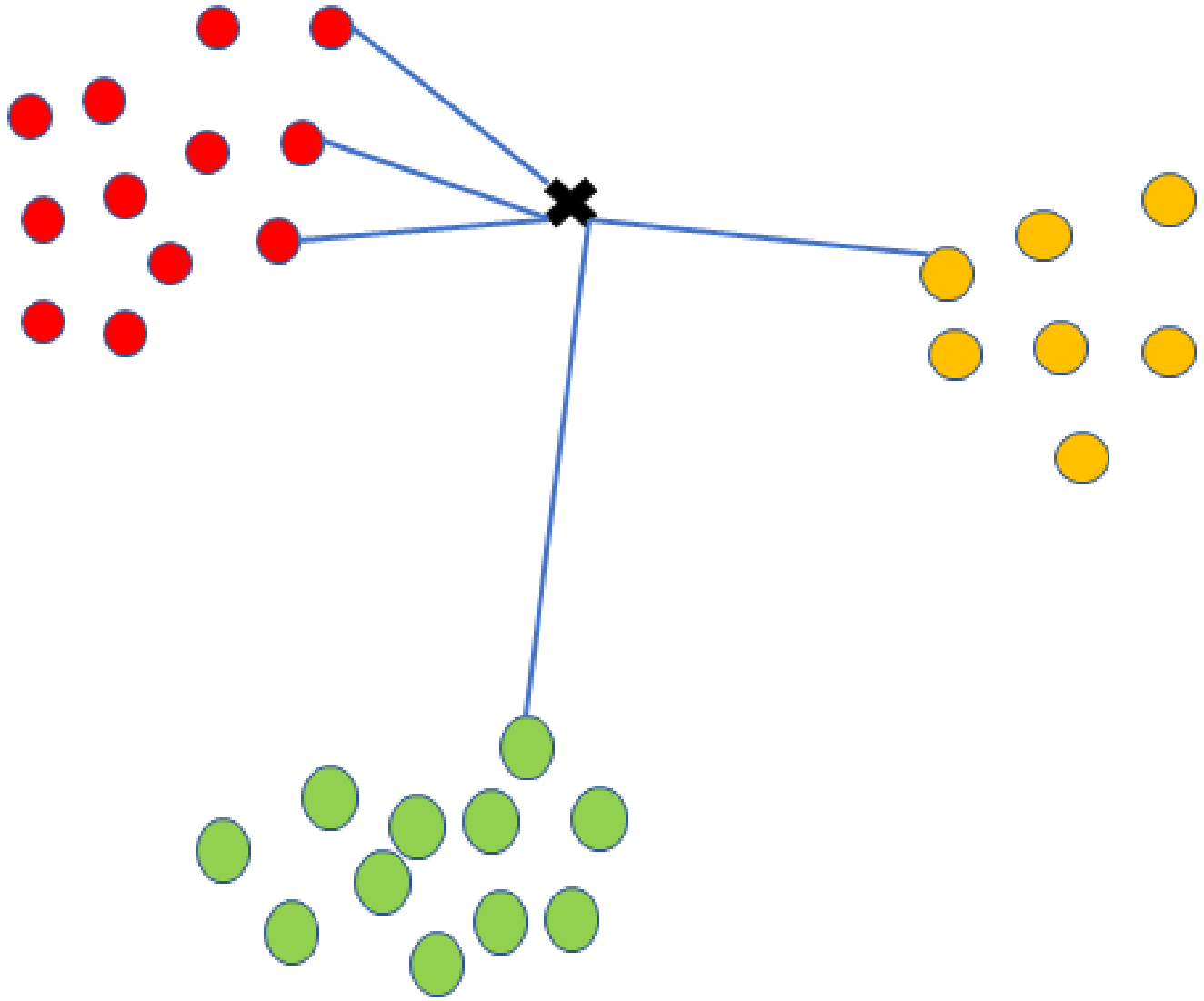
K=3

A Sample (K-NN)

K=5



Challenge at the Boundaries



Advantages

- Simplicity
- Flexibility
- Non-Parametric
- Easy Adaptation

Disadvantages

- Computational Cost
- Memory Usage
- Imbalanced Datasets
- Selecting the Optimum K Value
- Need for Feature Scaling

X_1	X_2	D
2	3	**A**
5	4	**B**
0	5	**A**
-3	0	**B**

Manual Calculations

By using Manhattan distance, find the class of (-2, 2) point.

X_1	X_2	D
2	3	**A**
5	4	**B**
0	5	**A**
-3	0	**B**

Manual Calculations

By using Manhattan distance, find the class of (-2, 2) point.

At first, calculate all distances:

d_1=|-2-2|+|2-3|=**5** d_2=|-2-5|+|2-4|=**9**

d_3=|-2-0|+|2-5|=**5** d_4=|-2+3|+|2-0|=**3**

For K=1… Choose its class as B.

For K=3… By majority among classes (A, A, B), choose its class as A.

Machine Learning

2. week

Thanks for watching