K Nearest Neighbor (KNN)

Khundrakpam Veeshel Singh
NIELIT Imphal

Outlines

- Metric
- Distance
- What is KNN
- How it works
- Pros & cons

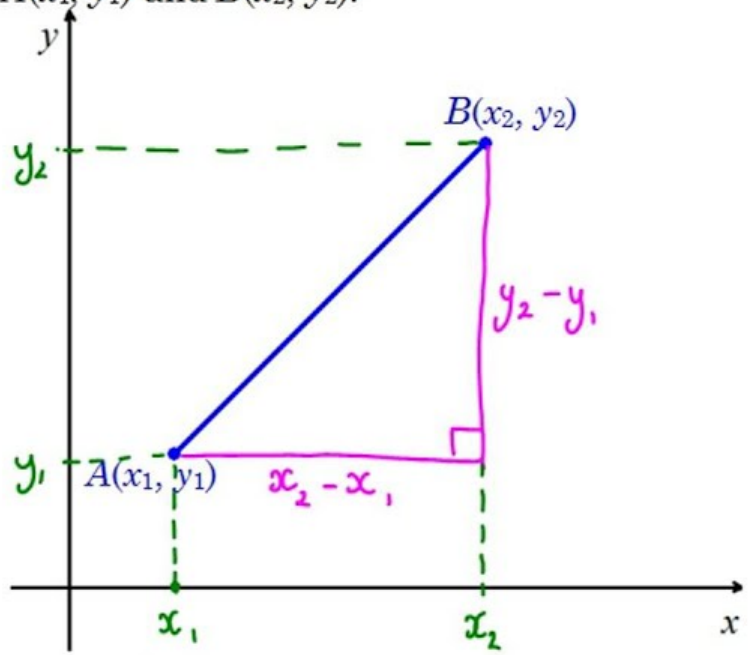
Metric

- Any measures that satisfies following properties are called metrics
  - Positivity, $d(x,y) \leq 0 \; \forall \; x,y$
  - Symmetry, $d(x,y)=d(y,x) \; \forall \; x,y$
  - Triangle inequality, $d(x,z) \leq d(x,y) + d(y,z) \; \forall \; x,y,z$

  Example distance

Distance

Let's consider this more generally by finding the distance between the points $A(x_1, y_1)$ and $B(x_2, y_2)$:



$$AB^2 = (x_2 - x_1)^2 + (y_2 - y_1)^2$$

The distance between two points $A(x_1, y_1)$ and $B(x_2, y_2)$ is given by:

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

Distance

- Minkowski distance:

$$d(x, y) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{\frac{1}{r}}$$

Distance

- When *r =1:* Manhattan distance or L1 Norm

$$d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k| \right)$$

- When *r=2*: Euclidean distance or L2 norm

$$d(x,y) = \left(\sum_{k=1}^{n} (x_k - y_k)^2 \right)^{\frac{1}{2}}$$

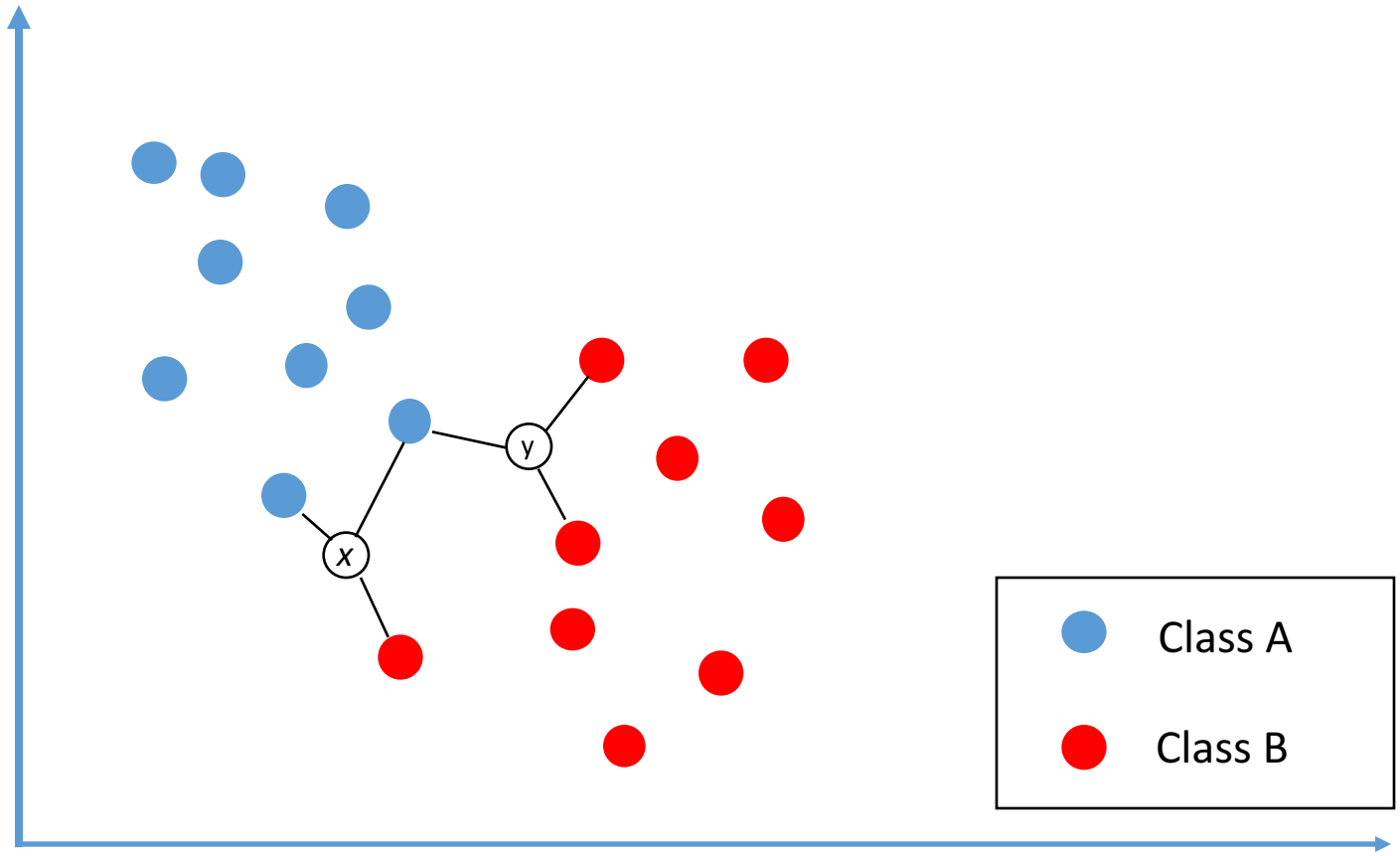
What is KNN

- The k-nearest neighbors algorithm, also known as KNN or k-NN, is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point.
- It is also call lazy learner
- It is instance based learning
- *“It looks like this, it feels like this,…, so it must be this”*

How KNN works

- Determine the K
- Calculate the distances of the input data to all the data points in the dataset.
- Select the closest k data points
- Majority voting among the k data points determine the output

KNN working



Pros & Cons

- Pros
  - Simple, easy to understand and easy to implement
  - No parameter
  - No training
  - Only one hyper parameter
  - Works on both classification and regression
  - Constantly evolves
- Cons
  - Slow
  - Need homogenous feature
  - Curse of dimensionality
  - Difficult to choose the optimal value of k
  - Noise or outlier sensitivity
  - Difficult to handle missing value

Some points to noted

- Value of k must not be multiple of no. of class
- Must be homogenous feature
- Dataset must be treated with dimensionality reduction and noise removal