DEPARTMENT OF COMPUTER SCIENCE
INSTITUTE OF MANAGEMENT AND RESEARCH, JALGAON

Name Take Pratiksha Raju
Expt. Title Implement simple KNN using Euclidean distance in Python.
Class _____ Batch B-4 Performed on _____
Roll No. 130 Expt. No. 05 Submitted on _____
Remarks _____ Returned on _____

* K- Nearest Neighbour (KNN) algorithm

K- nearest Neighbour is one of the simplest machine learning algorithm based on supervised learning technique. K-NN algorithm stores all the available data and classified new data point based on the similarity. This means a when new data appear then it can easily classified into a well suite category by using K-NN algorithm.

K-NN is a Non-parametric algorithm which means it does not make any assumption on underlying data. It is lazy learning algorithm where all assumption is deferred until classification

* Algorithm.

The KNN ~~algorithm~~ classification is performed using the following four step.

- compute the distance metric between the test data point and all the labeled data points.
- ordered the labeled data points in the increasing order of this distance metric
- select the top k labelled data points and look at the class labels.
- find the class label that the majority of these k labeled data points have and assign it to the test data point.

* Distance calculation formula:-

1) Euclidean Distance:-

It is generally used to find the distance beti two real-valued vectors.

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

2) Manhatten Distance:-

This is the simplest any or technique to calculate the distance between two point of ten called Taxicab distance or city Block distance.

Manhatten Distance = sum for 1 to N

sum $|| x1_i - x2_i ||$.

3) Hamming distance:-

The hamming distance is mostly used in text processing or having boolean vector. Boolean vector mean the data is in the form binary digits 0 and 1

Hamming distance $(x_1, x_2) =$ .....

4) Minkowski Distance:-

Minkowski distance is the generalization form of the euclidean and Manhatten Distance.

$$||x_1 - x_2|| = \left(\sum_{i=1}^{d} / x_{1i} - x_{2i} /^p\right)^{1/p}$$

DEPARTMENT OF COMPUTER SCIENCE
INSTITUTE OF MANAGEMENT AND RESEARCH, JALGAON

Name Take Pratiksha Raju
Expt. Title
Class FYMCA Batch B-4 Performed on
Roll No. 170 Expt. No. 05 Submitted on
Remarks Returned on

* Formula of information Gain

① Entropy :- $-p \log_2 p - q \log_2 q$

To build is a decision tree, we need to calculate two type of entropy using frequency table.

ⓐ Entropy using the frequency table one attribute

$$E(S) = \sum_{i=1}^{c} -pi \log_2 pi$$

ⓑ Entropy using the frequency table of two attribute

$$E(T, X) = \sum_{c \in X} P(c) E(c)$$

② $Gini = 1 \sum_{i=1}^{n} (Pi)^2$

③ Information Gain = 1 - Entropy.