**KNN: Hadoop and Map Reduce**

**SUBHAM SAHU**

**15BCB0105**

kNN is a non-parametric lazy learning algorithm. Being a non-parametric algorithm it does not make any assumptions on the underlying data distribution. This is a major advantage because majority of the practical data does not obey theoretical assumptions made and this is where non-parametric algorithms like kNN come to the rescue. kNN is also a lazy algorithm this implies that it does not use the training data points to do any generalization. So, the training phase is pretty fast. Lack of generalization means that kNN keeps all the training data. kNN makes decision based on the entire training data set.

The k-Nearest Neighbor Algorithm finds applications in some of the fascinating fields like Nearest Neighbor based Content Retrieval, Gene Expressions, Protein-Protein interaction and 3-D Structure predictions are to name a few.

* ***Requisites for k-Nearest Neighbor Algorithm***

kNN assumes that the data is in a feature space. More exactly, the data points are in a metric space.

The data can be scalars or possibly even multidimensional vectors. Since the points are in feature space, they have a notion of distance – This need not necessarily be Euclidean distance although

it is one of the commonly used methods.

The kNN uses training data as reference to classify the new data points collectively called testing dataset

Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either + or – (for positive or negative classes). But kNN, can work equally well with arbitrary number of classes.

We are also given a single number ‘k’. This number decides how many neighbors (where neighbors is defined based on the distance metric) influence the classification. This is usually an odd number if the number of classes is 2. If k=1, then the algorithm is simply called the nearest neighbor algorithm

* **The k-Nearest Neighbor Algorithm involves two phases.**

The Training Phase

The Testing Phase

***1) The Training Phase***

kNN Algorithm does not explicitly require any training phase for the data to be classified. The training phase usually involves storing the data vector co-ordinates along with the class label. The class label in general is used as an identifier for the data vector. This is used to classify data vectors during the testing phase

***2) The Testing Phase***

Given data points for testing, our aim is to find the class label for the new point. The algorithm is discussed for k=1 or 1 Nearest Neighbor rule and then extended for k=k or k- Nearest Neighbor rule.

**THE HADOOP FRAMEWORK**

Hadoop is an open source software framework that can run large data-intensive, distributed applications. Hadoop comes with its own file system called the Hadoop Distributed File System (HDFS) and a strong infrastructural support of managing and processing huge petabytes of data.

A HDFS cluster consists of one unique server known as the Namenode that manages the namespace of the file system, determined the mapping of blocks to Datanodes, and regulates file access. Each node in the HDFS cluster is a Datanode that has the task of managing the storage it holds. It is these

Datanodes that serve the client’s read/write requests and perform block operations, creation, deletion and replication instructions from the Namenode.

**ALGORITHMIC DESIGN OF THE K-NEAREST NEIGHBOR (KNN) TECHNIQUE IN THE MAPREDUCE PARADIGM**





