High Performance Computing : Assignment 4

Parallel Implementation of the k Nearest Neighbors Classifier

Oral questions

1. What is K in the K nearest neighbors algorithm

2. How is KNN algorithm implemented?

First, the distance between the new point and each training point is calculated by using Euclidean distance. Sort the data based on distance. The closest k data points are selected.

if value of k is 3 then first 3 rows are selected.

3. Write kNN Algorithm for Manual Implementation

4. What is the advantage of K nearest neighbor method?

   Advantages of KNN

   1. No Training Period: KNN is called **Lazy Learner (Instance based learning)**. It does not learn anything in the training period. It does not derive any discriminative function from the training data. In other words, there is no training period for it. It stores the training dataset and learns from it only at the time of making real time predictions. This makes the KNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression etc.

   2. Since the KNN algorithm requires no training before making predictions, **new data can be added seamlessly** which will not impact the accuracy of the algorithm.

   3. KNN is very **easy to implement**. There are only two parameters required to implement KNN i.e. the value of K and the distance function (e.g. Euclidean or Manhattan etc.)

5. What are some applications of KNN?

   Text mining.
   Agriculture.
   Finance.
   Medical.
   Facial recognition.
   Recommendation systems (Amazon, Hulu, Netflix, etc)

6. What is the best way to choose K in Knn?

7. The optimal K value usually found is **the square root of N**, where N is the total number of samples. Use an error plot or accuracy plot to find the most favorable K value. KNN performs well with multi-label classes, but you must be aware of the outliers

8. Compare sequential and parallel implementation

   Sequential o(n) as all eucledian distance calculated serially

   Parallel best – o(1)

9. How do you draw a KNN decision boundary?

   Perpendicular bisector of nearest neighbors of each class/cluster

10. How can I improve my Knn accuracy?

    -optimal k value

    - knn is sensitive to outliers so scale the data (fit & transform)

11. Is Knn supervised?

    yes

12. What is cross validation in Knn?

    The goal of cross-validation is **to estimate the expected level of fit of a model to a data set that is independent of the data that were used to train the model**. It can be used to estimate any quantitative measure of fit that is appropriate for the data and model.

13. How do you determine the value of K in K means?

    The Elbow Method

    The Silhouette Method

Practice problem

- Implement Knn using machine learning in Python?
- Implement parallel KNN using cluster of Raspberry Pi
- Check the performance of this program by varying number of nodes in a cluster and plot the graph.
- Implement Parallel KNN for large data set

1. Implement KNN by varying the distance formula.(Hamming Distance
2. Euclidean Distance
3. Manhattan Distance

-)