Assignment 3

Kapil Kumar(2014CS50736), Mahdihusain Momin(2014CS50288)

April 17, 2018

1

1.1 a

Sequential algorithm maintains a max heap to store top-k nearest neighbour. For each query point, it scans all the data points and add to heap if its distance to query point is less than top element in max heap.

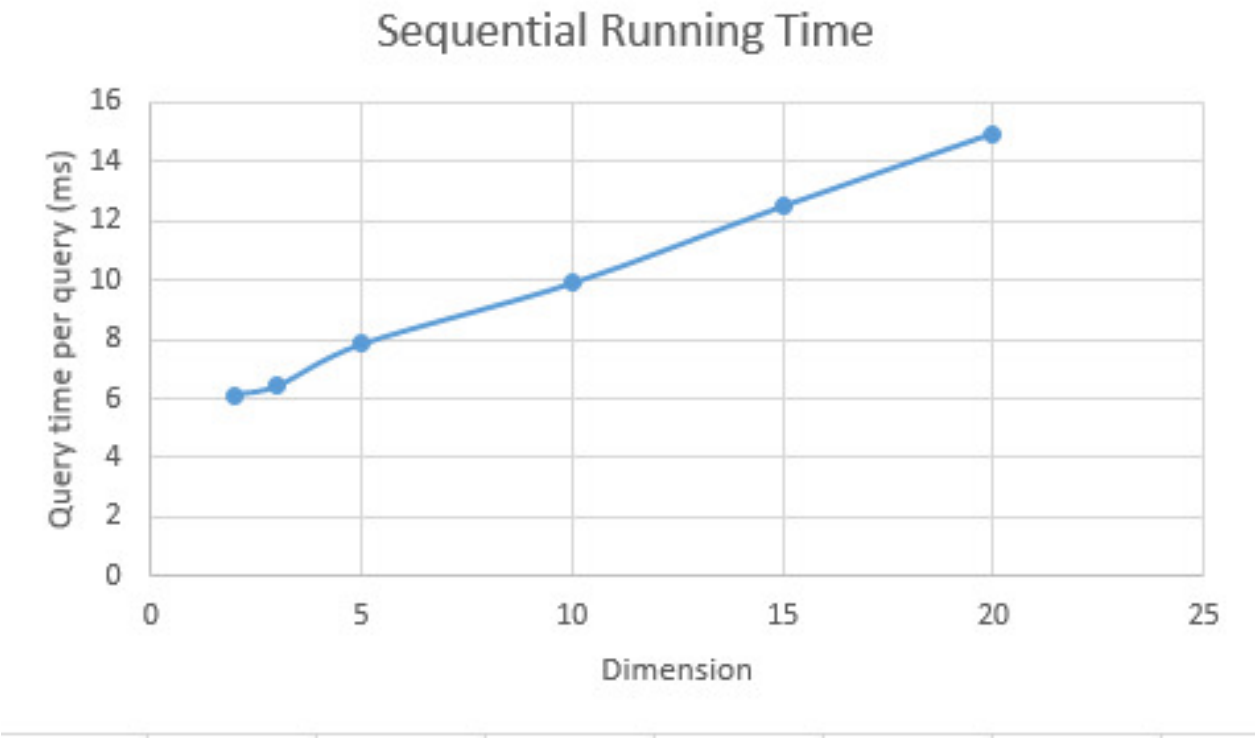

Figure 1: Dimension vs Avg. query running time graph for sequential algorithm

Best first search algorithm maintains a kd-tree data structure of data points. It prunes out some branches based on minimum distance between query point and minimum bounding rectangle(MBR). Here also, we maintain a max heap to store top-k nearest neighbours.

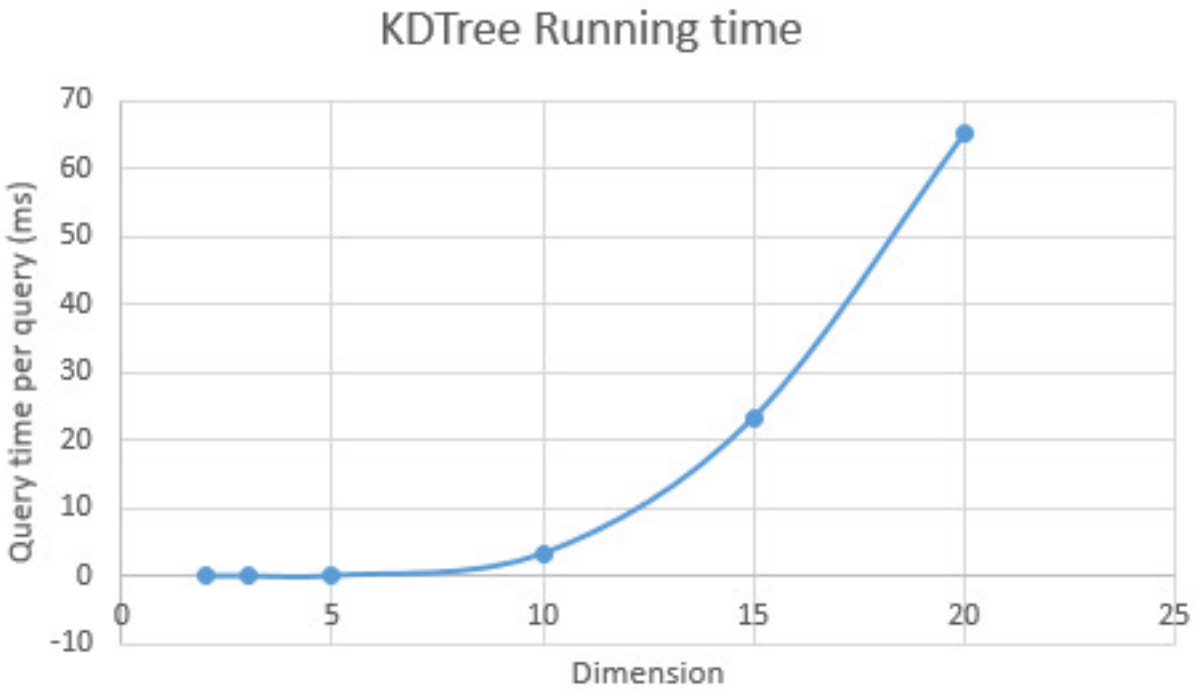

Figure 2: Dimension vs Avg. query running time graph for best first search in kd tree

1.2 b

Dimension	Avg. 2nd Neighbor Distance	Avg. 100th Neighbor Distance
2	0.00234739	0.0181686
3	0.015788	0.0644402
5	0.0821552	0.195577
10	0.333687	0.536077
15	0.586365	0.817753
20	0.831737	1.08538

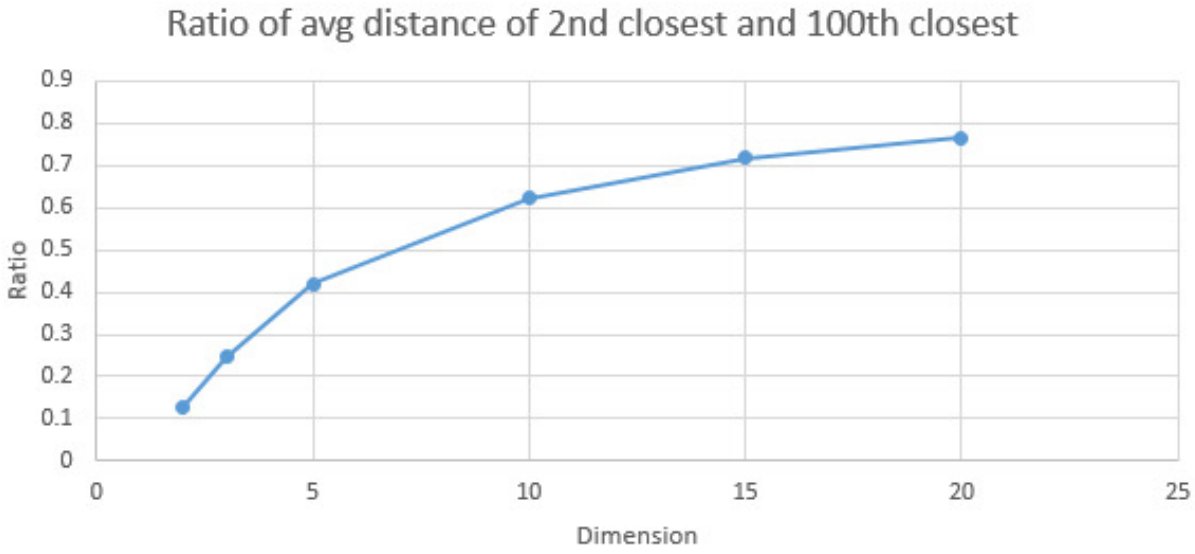

Figure 3: Ratio of Avg distance of 2nd nearest neighbour with 100th nearest neighbour

1.3 c

1.3.1 Reason for Running Time of Sequential Algorithm

Running time of calculating euclidean distance of two points increases linearly with increase in dimension. So, when total no. of data points are equal, avg running time increases linearly with dimension in figure 1.

1.3.2 Reason for Running Time of Best First Search in KD-tree

As you can see in figure 2, running time of best first search in kd-tree is actually better in lower dimension than higher dimensions. In higher dimensional spaces, because of curse of dimensionality algorithm visits many more branches than in lower dimensional spaces. The minimum distance(difference in splitting dimension) between query point and Minimum Bounding Rectangle(MBR) becomes comparably small with large dimension than actual distance with any point in MBR. Thus, there is very less pruning in case of larger dimension.

1.3.3 Reason for Ratio of Avg. distance of 2nd NN with 100 NN Graph

As you can see in figure 3, ratio of avg. distance of 2nd nearest neighbour and 100th nearest neighbour increases with increase in dimension. As dimension increases, space has increased but number of data points remain same. So, avg distance of 2nd nearest neighbour and 100th nearest neighbour will increase but the effect of dimension will be more on 2nd nearest neighbour because 100th nearest neighbour was already far from the point.

1.4 d

Competition.