Dylan Mohsen

DS210

HW 9 Report

Data Generation (generate_data): Random points are generated with x-coordinates in the range of -100,000,000 to 100,000,000 and labels of either 0 or 1.

Data Reading (read_data): The generated data points are read from a file, parsed, and stored in a vector for further processing.

Finding the Best Threshold (find_best_threshold): The program iterates over each point in the dataset, considering the x-coordinate of the current point as a potential threshold. For each potential threshold, it predicts the label for all points based on whether their x-coordinate is greater than or equal to the threshold (predicted label 1) or less than the threshold (predicted label 0). The accuracy of these predictions is compared to the actual labels, and the threshold with the highest accuracy is selected as the best threshold.

The complexity of the solution is determined by the find_best_threshold function. This function has a nested loop, so for each point in the outer loop, it iterates over all points again in the inner loop to calculate the accuracy of classification, thus the overall time complexity is O(n^2) for n points.

Output:

if x >= -87489081

Predicted label is 1

else

Predicted label is 0

accuracy: 0.56

I spent about 7 hours on this homework, and ran into many issues coding that made it take longer and frustrating.