

Lab 9 – NIST 19

Abhisek Mohanty
Abhishek Nigam

Ideas / Approach

■ First Approach

- Cluster data into 20x20x20 bins, based on certain thresholds for each variable.
 - Bin sizes constant at 20x20x20
 - For each dataset, different thresholds for speed, flow and occupancy based on data extrema.
 - **Intuition behind this idea was to divide the data equally using threshold which would result in a variable number of vectors in each bin.**

■ Second Approach

- Bins not of similar sizes in the above approach. E.g. 4.5million (0,0,0) values
- Cluster data into equal sizes instead of a threshold.
- **Intuition behind this idea was to divide data equally in bin making it easier to identify outliers.**

Implementation

- Cluster data on X axis based on data sorted by flow column.
- For each bin, extend the bin data onto the Y-plane, based on data sorted by speed column
- For each bin on X-Y plane, extend onto Z-axis, based on data sorted by occupancy values

Results and Analysis

- Distribution of densities more extreme in approach 2 compared to 1, lowest and highest densities are concentrated towards the ends.
- Approach 1 works better at isolating the possible outliers compared to approach 2, which might club some good data with outliers.
- An extension of Approach 1 would be to use Kmeans to cluster the data into 20x20x20 bins and then calculate the probability.

Self-evaluation and improvement of the final results

- We tried two approaches which portrayed different result values for probability distribution. But in our analysis we found that the result more or less corresponded to the same outliers.
- We gained invaluable insight into outlier detection. We also further tried K-Mean Clustering for the improvement of our final results. Due to performance and time constraints we were not able to show the results. It is running on our machine and we will be happy to share results when it completes.

Questions





Herbert Wertheim
College of Engineering
UNIVERSITY *of* FLORIDA

POWERING THE NEW ENGINEER TO TRANSFORM THE FUTURE