**Preprocessing:**

The raw data read from the csv file was preprocessed by making a row with length equal to the number of weeks and in each training example it gives for the specific week the day when the mall was visited or 0 if it wasn’t. it follows the same numerical distribution of weekdays mentioned in the file. The data in the table are shuffled then divided into training and test sets. Usually we divide into 7:3 ratio. However, since the dataset is huge we can keep 5:1 ratio.

**Algorithms:**

The used algorithm is Knn with hamming distance as a metric since we are looking for the closest items that is identical to the test sample as much as possible. So the distance between [4,1,0,2],[5,7,3,1], and [4,7,0,3] are 2 and 3 respectively divided by the total # of weeks.

To pick the best K, cross validation test was applied with 10-fold sets. It gave a best accuracy of k =1

The overall accuracy was around 41% which is pretty fine knowing that the raw data doesn’t have enough features to classify upon (I also tried clustering algorithms but didn’t return a good accuracy)

The evaluation metric used here is the accuracy of classification.

**Performance**

Since Knn is a lazy algorithm, it will take a lot of time to examine all the test inputs. However, this code can be easily parallelized.

The code runs on python 3.6