MSDS 7333 Spring 2021: Case Study 01

Real Time Location Systems

Sachin Chavan,Tazeb Abera,Gautam Kapila,Sandesh Ojha

2021 January 19

Introduction

Real time location systems are used to locate objects and people in real time. For enterprises from different industries it is crucial to locate its assets which in turn helps increase in performance and improve services. Global positioning systems are most popular nowadays. One of the example from our daily lives is when we order food or cab online we can watch real time location of cab or delivery person and we can see estimated time of receive services at our end. Global positioning systems which uses satellite signals to track objects works only outdoor. They don't work inside buildings.

With widespread use of Wireless technologies, we now have indoor positioning systems as well. These systems works very well and efficiently inside buildings. There are various ways of implementing Indoor positioning systems like Infrared systems, Proximity based systems, Acoustic System and WiFi based systems to name a few. Indoor positioning systems helps companies to track people and different assets in real time which they can use to improve productivity and services which in turn helps to increase profits.The purpose of such systems is to monitor movement of its people and assets in real-time, thereby reducing time spent in finding assets. The main idea of tracking things is to analyze productivity, improve services and increase efficiency which in turn helps in profits.

Business Understanding

The business need that this work addresses is ability to locate objects and people on a given floor of a work or factory environment. This helps quantifying the efficiency of movement of assets, and can enable companies to put metrics around it that can be tracked year on year, leading to cost savings. E.g. in aerospace industry, where aircraft assembly can be a very complex process, wherein, deployment of this system could help uncover (a) bottlenecks in production line, (b) over all asset tracking and management, and (c) improve prediction in completion of assembly task.

This case study evaluates WiFi based Real time Location System for an organization. The dataset provided for this case study contains one million measurements of signal strength recorded at six different stationary access points (WiFi routers). These signal strengths are measured between handheld device such as cellular phone, laptops and all six access points. The goal of this study is to build a model using this dataset to detect the location of the device as a function of strength of the signal between handheld device and each access point and use this model to predict the location of the device based on the strength of the signal between device and each access points.

Layout of the building floorplan is depicted in Fig. 1

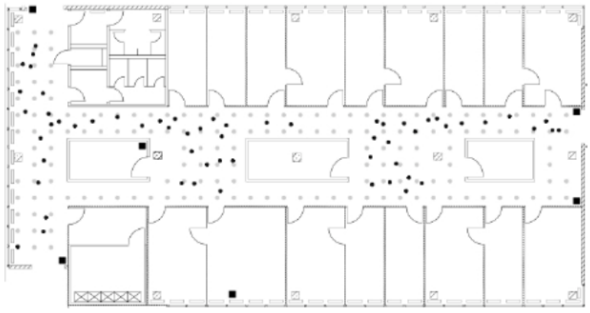

Figure 1: Building Floorplan

As shown in the Fig.1 Six stationary access points are denoted by black square dots. Signal strength between handheld device and each access points was measured at 166 different locations with 8 different angles (0, 45,90,135 and so on) on this floor marked by grey dots. All grey dots are spaced one meter apart. Online measurements were recorded randomly selected points indicated with round black dots.

There are numerous algorithms available to estimate location of the device from strength of the signal between device and each access point. This is classification problem and in this case study simple k-Nearest Neighbor (kNN) algorithm will be used as a classifier to build a model. Since training dataset contains signal strength between device and access points at 166 different locations, Idea is for every new device on the floor with known signal strength find k nearest neighbors with similar signal strength at the known locations in training data by calculating **Euclidean distance** between two sets of signals as follows.

$$\sqrt{\sum_{i=1}^{6}(S_i * - S_i)^2}$$

Where,

S_i* = Single strength between Access point and new device

S_i = Single strength between Access point and specified position in the training data

Objective

Build a model using offline data set to predict location of the devices in online dataset.

Two methods that shall be used for this case study are:

1. kNN
2. Weighted kNN

Data Evaluation / Engineering

In order to build Indoor Positioning System two datasets have been made available.

1. offline.final.trace.txt

   This data will be used to train the model

2. online.final.trace.txt

   This is test dataset.

Both files are variable length files made up of following fields:

Field Name	Field Description
t	timestamp in milliseconds since midnight, January 1, 1970 UTC
id	MAC address of the scanning device
pos	the physical coordinate of the scanning device
degree	orientation of the user carrying the scanning device in degrees
mac	MAC address of a responding peer (e.g., an access point or a device in adhoc mode) with the corresponding values for signal strength in dBm (Decibel-milliwatts), the channel frequency and its mode (access point = 3, device in adhoc mode = 1)
signal	Signal Strength in DbM

Both file are structured in specific format using more than on delimiters.Files contain fields related to scanning device and access points. Since data is not in tabular format some string manipulations has been performed on the dataset to convert it into tabular format.

Field mapping between text file and DataFrame as follows:

Field ID	Total DF Fields	New fields created in dataframe
t	1	time
id	1	scanMac
pos- These are comma separated fields x,y,z coordinates	3	posX,posY,posZ
degree	1	orientation
MAC id of access point	1	mac
MAC is of access points are followed by three fields Signal strength,channel, access point type	3	signal,channel and type

Struture and summary of dataframe after mapping all fields from input file is as follows:

DataFrame Structure:

```
## 'data.frame':    1181628 obs. of  10 variables:
##  $ time       : num  1.14e+12 1.14e+12 ...
##  $ scanMac    : chr  "00:02:2D:21:0F:33" "00:02:2D:21:0F:33" ...
##  $ posX       : num  0 0 0 0 0 ...
```

```
##  $ posY       : num  0 0 0 0 0 ...
##  $ posZ       : num  0 0 0 0 0 ...
##  $ orientation: num  0 0 0 0 0 ...
##  $ mac        : chr  "00:14:bf:b1:97:8a" "00:14:bf:b1:97:90" ...
##  $ signal     : num  -38 -56 -53 -65 -65 ...
##  $ channel    : chr  "2437000000" "2427000000" ...
##  $ type       : chr  "3" "3" ...
```

Dataframe Summary:

```
##       time             scanMac               posX            posY
##  Min.   :1.140e+12   Length:1181628     Min.   : 0.00   Min.   : 0.000
##  1st Qu.:1.140e+12   Class :character   1st Qu.: 2.00   1st Qu.: 3.000
##  Median :1.140e+12   Mode  :character   Median :12.00   Median : 6.000
##  Mean   :1.140e+12                      Mean   :13.73   Mean   : 5.876
##  3rd Qu.:1.140e+12                      3rd Qu.:23.00   3rd Qu.: 8.000
##  Max.   :1.142e+12                      Max.   :33.00   Max.   :13.000
##       posZ    orientation        mac                signal
##  Min.   :0   Min.   :  0.0   Length:1181628     Min.   :-99.00
##  1st Qu.:0   1st Qu.: 90.0   Class :character   1st Qu.:-73.00
##  Median :0   Median :180.0   Mode  :character   Median :-62.00
##  Mean   :0   Mean   :167.2                      Mean   :-63.85
##  3rd Qu.:0   3rd Qu.:270.0                      3rd Qu.:-55.00
##  Max.   :0   Max.   :359.9                      Max.   :-25.00
##    channel             type
##  Length:1181628     Length:1181628
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
```

Following changes were made based on analysis from Descriptive statistics :

- Removal of the Z position because it is all zeros based on summary statistics.
- Making scan angles consistent throughout the dataset.
- Remove extraneous access points which are related to adhoc device type and those with fewer observations.
- Remove rows for type=1 as they are not access points.
- Drop column scanMac as there is only on scanning device. Removing this column won't affect analysis.

Updated Structure of DataFrame:

```
## 'data.frame':    6 obs. of  8 variables:
##  $ time   : POSIXt, format: "2006-02-11 01:31:58" "2006-02-11 01:31:58" ...
##  $ posX   : num  0 0 0 0 0 ...
##  $ posY   : num  0 0 0 0 0 ...
##  $ angle  : num  0 0 0 0 0 ...
##  $ mac    : chr  "00:14:bf:b1:97:8a" "00:14:bf:b1:97:90" ...
##  $ signal : num  -38 -56 -53 -65 -65 ...
##  $ rawTime: num  1.14e+12 1.14e+12 ...
##  $ channel: chr  "2437000000" "2427000000" ...
```

As we can see from above structure that mac now has only 7 levels. Which means that this dataset now removed all irrelevant data. But we have one extra accesspoint and we don't know which six are from the required floor of the building. Further analysis is required to confirm the same. Same is discussed in next section.

Updated DataFrame:

```
##                  time posX posY angle               mac signal      rawTime
## 1 2006-02-11 01:31:58    0    0     0 00:14:bf:b1:97:8a    -38 1.139643e+12
## 2 2006-02-11 01:31:58    0    0     0 00:14:bf:b1:97:90    -56 1.139643e+12
## 3 2006-02-11 01:31:58    0    0     0 00:0f:a3:39:e1:c0    -53 1.139643e+12
## 4 2006-02-11 01:31:58    0    0     0 00:14:bf:b1:97:8d    -65 1.139643e+12
## 5 2006-02-11 01:31:58    0    0     0 00:14:bf:b1:97:81    -65 1.139643e+12
## 6 2006-02-11 01:31:58    0    0     0 00:14:bf:3b:c7:c6    -66 1.139643e+12
```

This processed dataset will now be used for further analysis to find relationship between variables.

Modeling Preparations

1. Signal Strength

Figure 2a) shows signal strength as measuring device weakens with distance from access point while Figure 2b) shows standard deviation increases with average strength of the signal. These visualizations are strong indication of signal strength related to distance from the access point.This feature is going to help in modeling.

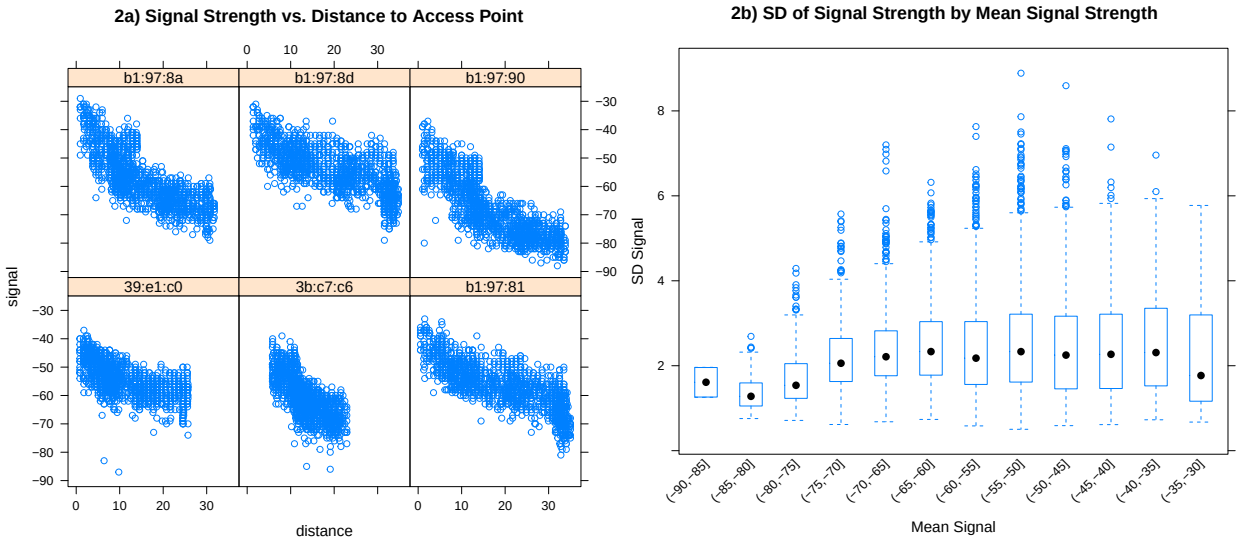

Figure 2: Signal Strength

2. Wireless Access Points

Initial analysis found that dataset contains 21 MAC addresses with two types of measurements adhoc (type 1) and access points (type 3). We know from the building floorplan we need to consider 6 fixed access points for the analysis,there records belonging to adhoc measurements are removed. This reduces number of MAC addresses from 21 to 12. As per documentation there are 5 Linksys routers and 1 Lancom router is placed in the building. Upon further analysis it is found that 5 MAC addresses belongs to Linksys and none belongs to Lancom. So based on number of measurements top 7 MAC addresses are kept for analysis. Since we have one extra MAC address in the dataset we need to test model with below configuration to find 6th access point that results in better location predictions based on signal strength.

MAC Address	Scenario 1 (C0)	Scenario 2 (CD)	Scenario 3 (COCD)
00:0f:a3:39:e1:c0	X		X
00:0f:a3:39:dd:cd		X	X
00:14:bf:b1:97:8a	X	X	X
00:14:bf:3b:c7:c6	X	X	X
00:14:bf:b1:97:90	X	X	X
00:14:bf:b1:97:8d	X	X	X
00:14:bf:b1:97:81	X	X	X

Constraints:

Access point locations are the same in training (offline) and test (online) dataset. If the locations change, then the signal strength – distance metric will change, and model has to be re-built.

Modeling Scenarios (original Case)

To determine which MAC address to keep, the below analysis is performed. Based on the output we need to determine whether to keep MAC address ending in **"CD"** only, the MAC address ending in **"C0"** only, or keep both in our dataset. KNN methodology is executed on all three configurations and we will use the error rates output of the KNN below to determine the configuration to keep. For each configuration K is selected from 1 to 20 and we will use the graph and the table below to determine the best fit.

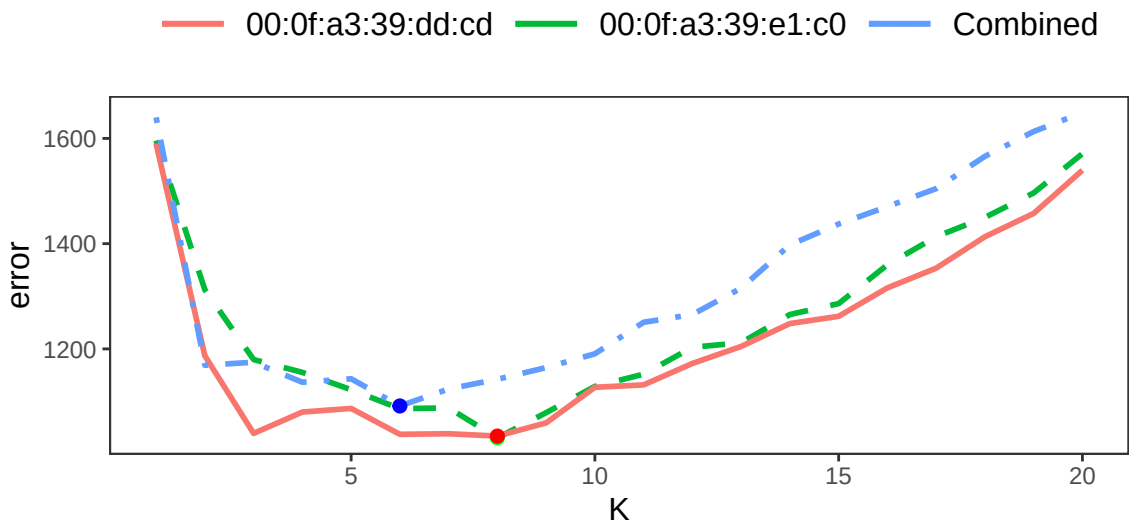

Figure 3: Simple Average KNN - Learning Curves for Each Scenario

Based on the combined elbow plot above we can see that for Simple Average KNN **"CD"** and **"C0"** has the exact K=8. For the combined K=6 looks to be the optimal solution.

Table 4: Simple Average Summary

Access Points	Best K	Error @ Best K
C0	8	1030.984
CD	8	1034.328
C0CD	6	1091.583

From the summary above the combination has the highest error. This would lead to not keeping both values in the mix. Given **"C0"** by itself has the lowest error this seems most likely option to address based on simple Average KNN analysis.

Modeling Scenario (extending Case)

Based on simple KNN we have determined that the dataset with **"C0"** only looks to be the best option. We will now perform a secondary analysis to verify the output from simple KNN using a weighted KNN. This methodology will allow for improved accuracy in predicting locations and the locations of closer distance may now have a larger impact to their neighbor groups then the once that are further away. We will also leverage cross fold validation as part of this weighted KNN. For this analysis we will run 11 fold validation and keep the same K values from 1 to 20.

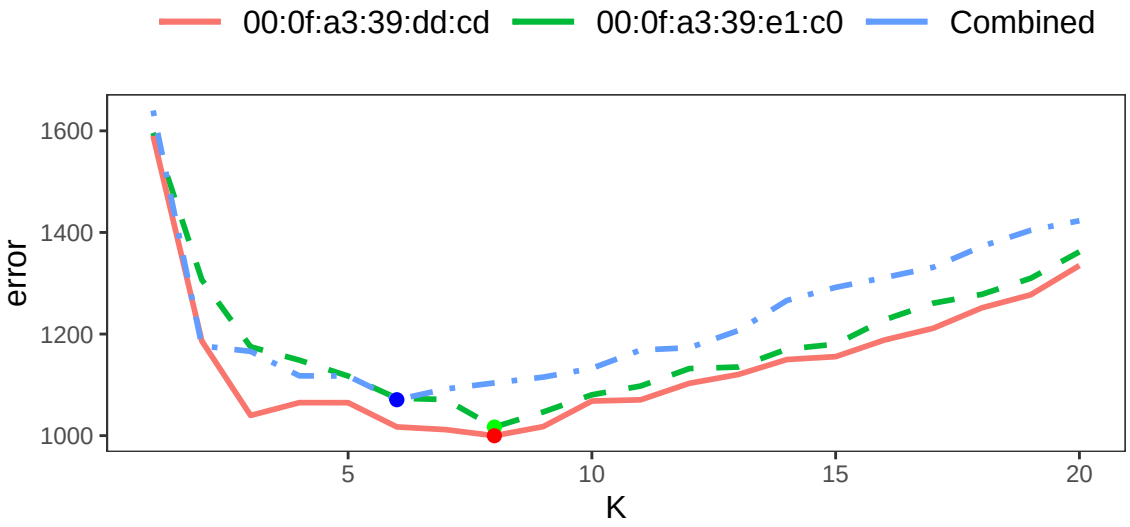

Figure 4: Weighted Average KNN - Learning Curves for Each Scenario

The K values remained the same for both Weighted and Simple KNN. Now, when we look at the weighted error below we can see that **"CD"** is slightly better than the other scenarios presented.

Table 5: Weighted Average Summary

Access Points	Best K	Error @ Best K
C0	8	1016.9812
CD	8	999.8401
C0CD	6	1070.7077

Best Scenario

Isolating just **"CD"** and looking at the elbow plt below shows that K is indeed 8 and it seems to be the optimal number of neighbors for this KNN analysis.

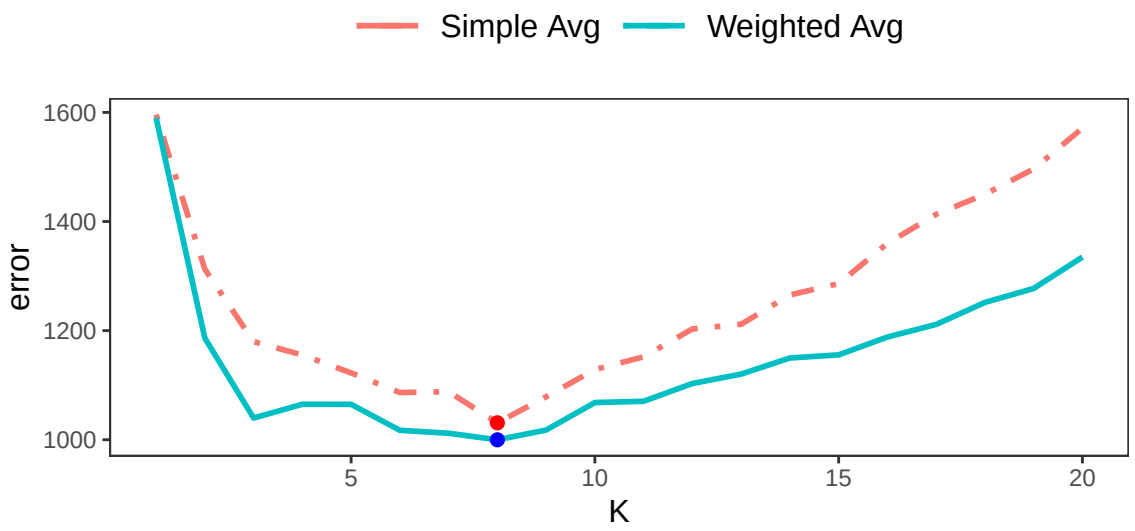

Figure 5: Best Scenario - Simple Avg (C0) Vs Weighted Avg KNN (CD)

The difference between predicted location and test location can be visualized as shown below. This allows determining the proximity of points that are together.

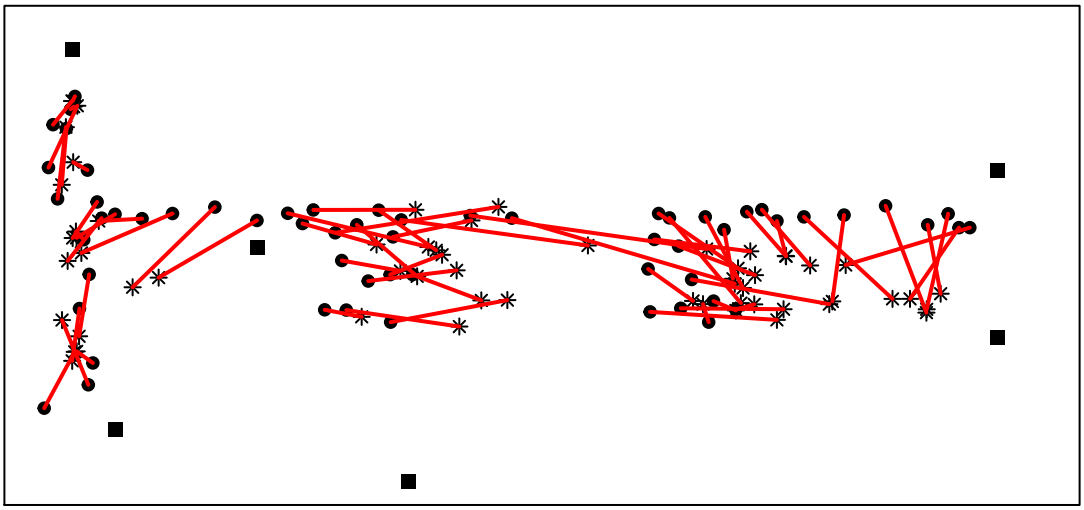

Figure 6: Floor Level Errors Using Weighted KNN with 8 Neighbors

Conclusion

Statistical indoor positioning system (IPS) for the test environment is developed using KNN algorithm. Three different access point combinations were analyzed, and the combination containing access point **00:0f:a3:39:e1:c0** provided best result as determined through minimum sum of squared error = **1030.984** for K = **8** using simple average method.

To explore generating better model than simple average approach above, weighted KNN approach is adopted. In weighted KNN approach, weights are setup to be inverse of distance metric. This ensures higher weights are provided to neighboring points, rather than far away point. Intuitively, signal strength is expected to be better correlated in neighboring points than far away points. This is also evident from signal strength vs distance relation. The weighted KNN approach appears to leverage above fact, and provide even lower sum of squared error metric **999.84** (**00:0f:a3:39:dd:cd**) for K=**8**, than simple average approach for any access point combination & K value, as noted previously.

One drawback of the approach is that it creates dependency on existing setup of WIFI access points. It is important that these are stationary, and their mac addresses are unique. Anytime a new WIFI router is installed, existing location is upgraded, or its location is changed, the repeat collection of offline training dataset needs to be undertaken, leading to overhead maintenance cost.

Alternative systems, e.g. use of Bluetooth beacons could also be considered for cost effectiveness, and improving accuracy of indoor positioning system.

References

[1] Deborah Nolan; Duncan Temple Lang. Data Science in R.Chapman and Hall/CRC, 2015.