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

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

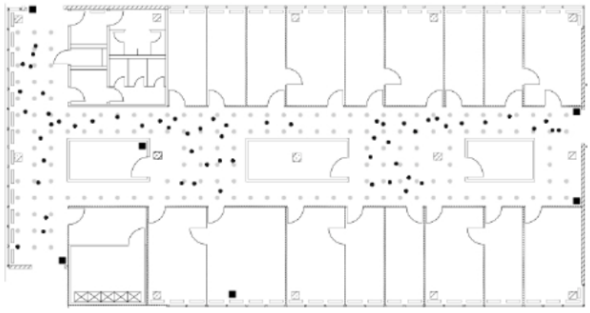

Figure 1: Building Floorplan

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 speficied 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:

```
## '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. Relationship between signal strength and distance

Signal strength at measuring device weakens with distance from Access Point. For the six access points in baseline analysis, the relationship is shown in Fig. 2

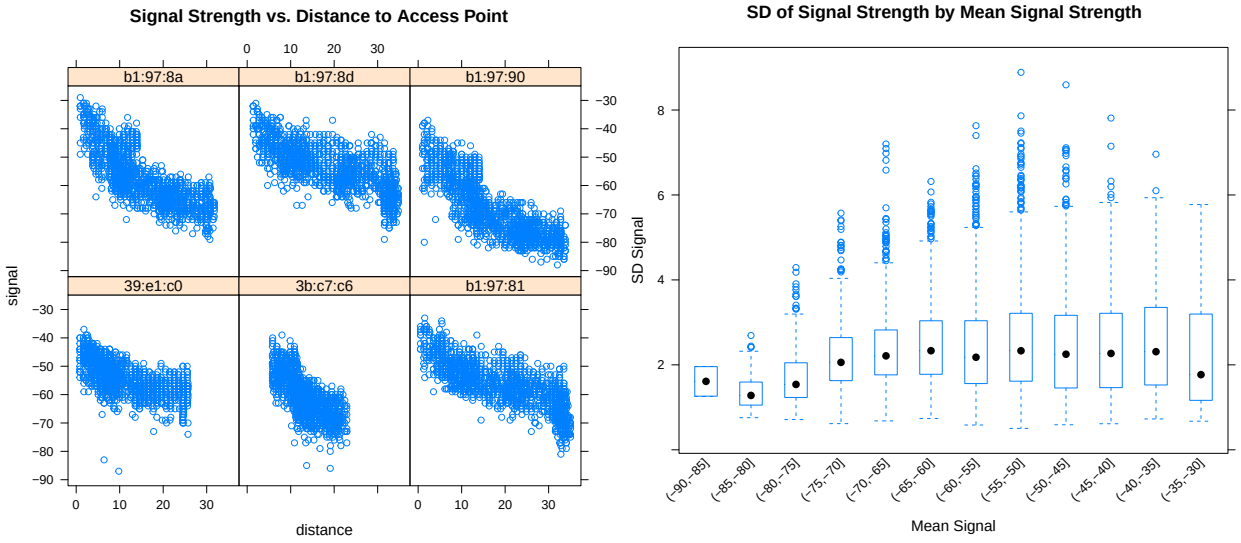

Figure 2: Signal Strength

2. Wireless access points are fixed

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.

Table 3: Formatted Dataset for KNN Use

posXY	posX	posY	00:0f:a3:39:e1:c0	00:14:bf:3b:c7:c6	00:14:bf:b1:97:81
0-0	0	0	-53.47602	-66.18943	-64.53751
0-1	0	1	-52.90871	-66.17718	-65.91699
0-10	0	10	-55.25828	-65.03161	-66.36639
0-11	0	11	-54.16614	-67.93829	-68.84551
0-12	0	12	-54.45000	-68.17850	-70.83332

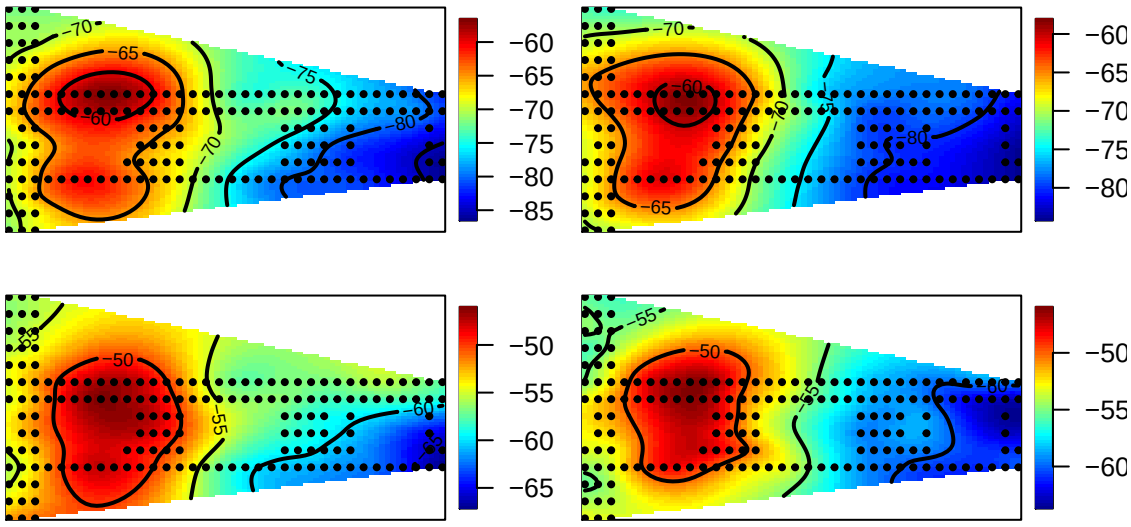

Figure 3: Signal Strength for Two Similar Access Points

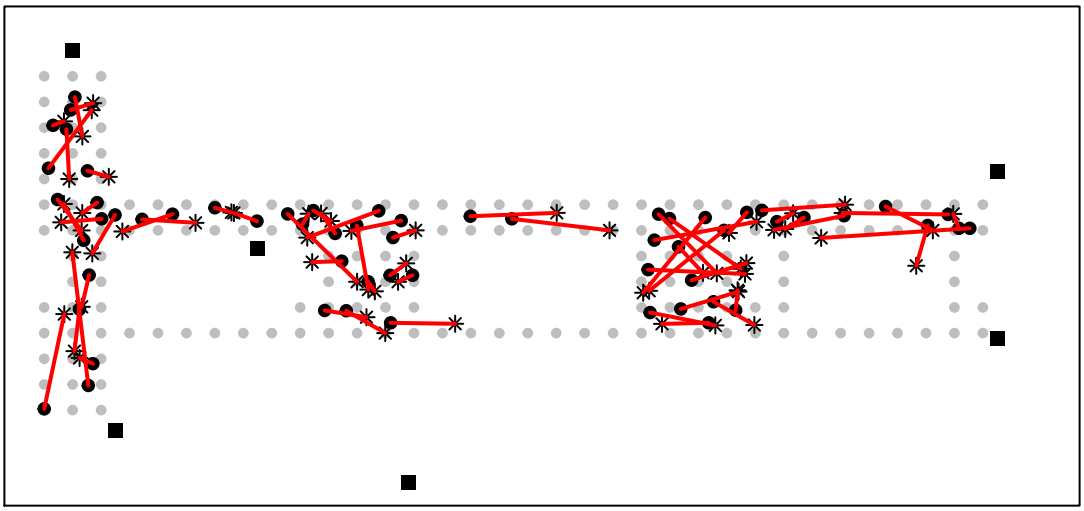

Figure 4: Floor Level Errors Using Weighted KNN with 6 Neighbors

Modeling Scenarios (original Case)

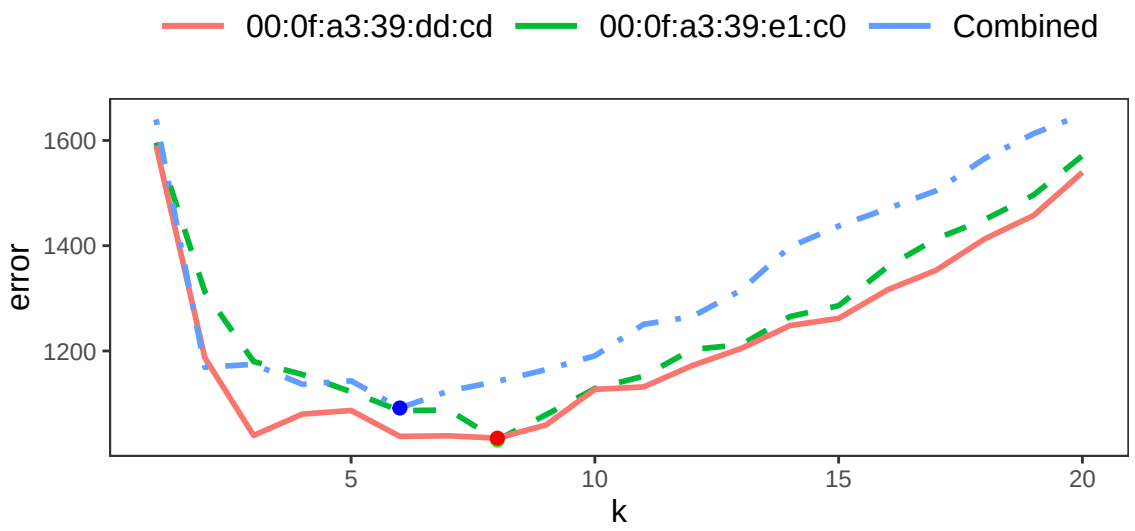

Figure 5: Simple Average kNN - Learning Curves for Each Scenario

Table 4: Simple Average Summary

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

Modeling Scenario (extending Case)

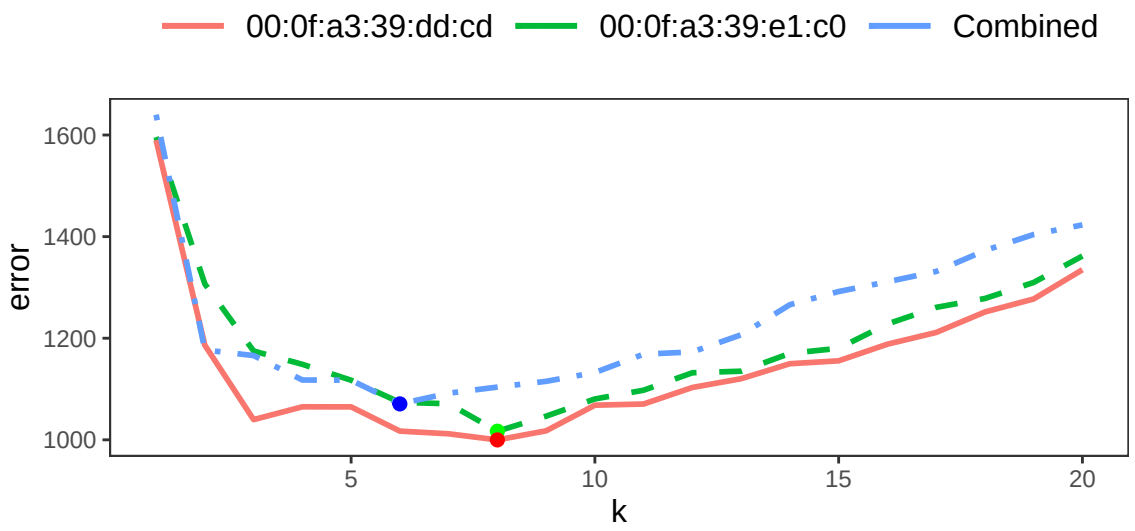

Figure 6: Weighted Average kNN - Learning Curves for Each Scenario

Table 5: Weighted Average Summary

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

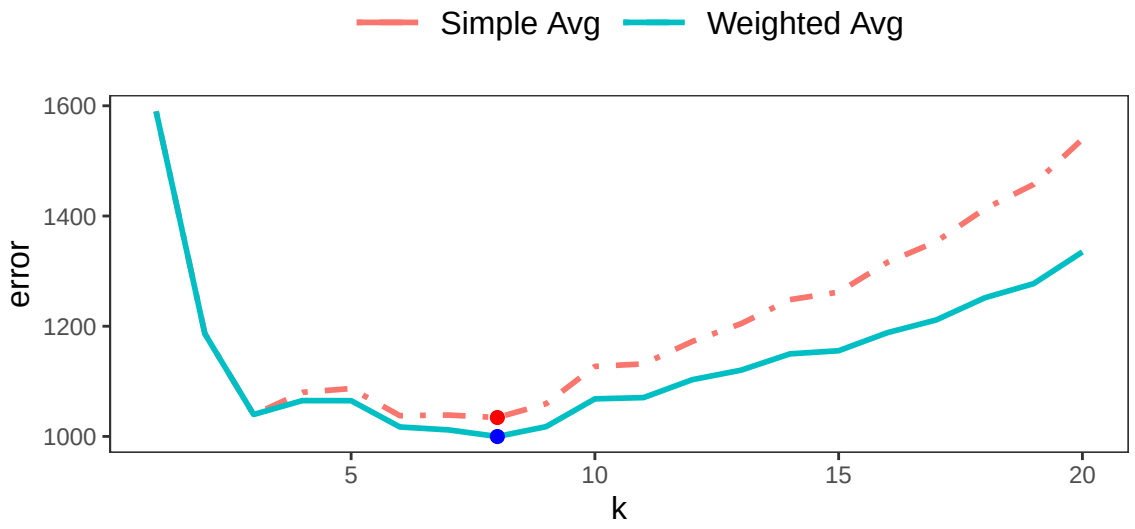

Figure 7: Best Scenario CD - Simple Avg Vs Weighted Avg kNN

Conclusions

References

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