Datasets

<http://rdatasciencecases.org/Data/offline.final.trace.txt>

http://rdatasciencecases.org/Data/online.final.trace.txt

**Homework 1:**

* Using the OFFLINE data and two different clustering methods predict the location of the ONLINE data set. (Offline – training, Online – live data)
  1. Describe how prepared data (explain steps to prepare for model, prep)
  2. Describe how you estimated error and found best fit assuming you can’t use the online data (describe the best fit for data, i.e. tables, charts w/analysis) (evaluating fit is subjective, subjectivity, what is important)
     + Describe the best fit for the data, If you cannot use Online data, then use KNN output and metrics; Look at the metrics related to Clustering: Silhouette Score – used to evaluate clustering models; Silhouette Score – Per scikit learn documentation, “The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). To clarify, b is the distance between a sample and the nearest cluster that the sample is not a part of. Note that Silhouette Coefficient is only defined if number of labels is 2 <= n\_labels <= n\_samples - 1.
     + Used KNN and different clustering method – Random Forest – or – DBSCAN – there are 4-5 measures to use to evaluate.
     + DBSCAN – BIC metric,
     + KNN – recommended because it is better at physical distance. Uses proximity
     + Gaussian mixture model – could be an option?
     + ROOT MEAN SQUARE ERROR –
     + \*\*READ THE KNN METRICS around 600ish-700ish\*\*
     + Could you use the Offline set only and cross validate, create data frame with x, y; use KNN to predict the x, y and compare; that would give you an accuracy score.

<https://www.rdocumentation.org/packages/cluster/versions/2.1.0/topics/silhouette>

There are 2 macIDs located … 2 were similar c0 and cd (used first, discared second). Use 6 ids to make prediction on outcome. Then SWAP (run 3 scenarios: id kept, id not kept, both)

3. Does one give better performance than the other? (compare c0 and cd, Q9 p 41)

4. What about using both? (comparison of c0 + cd against c0 only and cd-only

Step back, recommendations

5. What are drawbacks (if any of using this method to real-time locate and object) / What is the drawback (if any of using this method to real-time locate an object)?

6. Describe a method that may be improvement based on drawbacks / Describe a method that may be an improvement based on your perceived drawbacks.

k-NN with k =3 and k = 5 is not two different scenarios.

Round out Question 3 by Sat/Sun, few days to flesh out report

OTHER NOTES: Clustering Overview

NOTEBOOK provided gets data set up. Take textual data, parse, add back headers,

\*\*Orientation is important, Signal, MAC

* There are 1M rows, need to pare down

The code gets you to processed data.

NEXT STEPS in R Studio

* Wide data set – or cast data sets – long format, melted format. Now we make it wider to get the data structured, all records need to have the data.
* What does NA mean? Cannot do distance calc on row without data (i.e. w/NA, could impute values otherwise, but needs domain knowledge) – Call attention to the fact there are missing records
* Add posXY variable, captures x, y coordinate
* Then bring in online data.
* Make prediction on black dots, use gray dot info to figure out black dot (black dot likely at x, y)

OTHER QUESTIONS:

Focus on 7 macIDs from the text, 7 that have the most records; they opted to use those.

For angles, round to nearest 45 degrees…does that work? Is that best?

k-NN as clustering algorithm – k-Means is classification algorithm.

Identical – checks if the Offline data is the same as the OfflineReDo. After doing manipulation, is it the same?

Notes from Class on Case Study 1:

* Don’t recreate Chapter 1, don’t include things that are unnecessary
* Things to consider: Signal Strength, swap MAC ids (Q9), k-NN analysis, Weighted k-NN Analysis, Optimal k,
* Charts need explanation, either intro or conclusion, discuss insight; polished charts and visuals; data story narrative
* Focus on questions at hand, visualize/contextualize, then draw conclusions
* What are the novel things you’ve done

A Jupyter Notebook Example –

* Introduction (ask as well as answer, provide abstract overview of assignment)
* Methods/How you approached the problem (possibly data source, possibly the map)
* EDA / Exploratory data analysis. (the data came in unclean, then they layered in structure, what they accomplished in the early steps)
* Analysis (see below)
* Results (summarize)
* Conclusion
* References
* Appendix/Code (link out to GitHub reference)

A Word Doc Example –

* Summary
* Intro
* Methods
* Data Description
* KNN Model (what is it)
* Model Selection (summarize the optimal error rate)
* Visuals to compare/contract
* Conclusion/Summary
* Extensive Index
* References