Capstone

Memo

|  |  |
| --- | --- |
| To: | Mentor |
| From: | Yiching Huang |
| Date: | 2018/06/10 |
| Re: | Capstone – Locationing |

As you requested, this memo presents the result of the wifi location where I used the data analysis pipeline of what we’ve learned in this course from understanding the data, to using different models for training and testing the data and evaluating the results. I will detail the key takeaways that I gained from the process and potential business value from this analysis. I will separate the following discussion into different sections based on results.

**What is the data**

## The data that we used for the capstone report is the wifi location data. From attribute 1 to 520, it represents the wifi fingerprint value and ranges from -104 to 100. The value 100 is used if we cannot detect the wireless access points. Attributes 521 to 529 contain the different information inside and outside the building. In this task, our goal is to use wifi fingerprint to find out the inside location in order to help students to find the place they want to go inside the building since neither GPS nor Google Maps work well when we’re inside the building.

**Data Preprocessing**

We first imported the data to see if there’s any missing data. Next, in order to get the specific indoor location ID that we want to predict, we unite 4 attributes - Building ID, Floor, Space ID and Relative Position into one attribute – ID. Different from what we processed in R, we didn’t use the sampling method for our data. Instead, we chose to use the whole data for training and testing in the next steps. Thus, we simply identified WAP01-WAP520 are the independent variables and ID as our dependent variable.

**Data Training**

We used three different models for the data training, random forest, SVC and KNN. In our last data analysis process with R, we used C5.0 instead of SVC for this case. However, we do want to see if there’s a big different with different algorithms. After training the data with 3 different models and using the best parameters, we found that random forest can help us to generate the most optimal performance and achieve accuracy to 78%. Remember, the accuracy rate with the Random Forest method with R last time only achieved 66%. This big progress may result in we are using the whole population instead of the little sample from our data to proceed with the data analysis work and also because we use different coding in Python to find what might be the best parameter for the algorithm.

**Data Validation and Results**

We used the test data as our validation data and used the random forest method for predicting the ID location in test data. After compared with the ground truth, the accuracy rate achieves to 78.7% and the performance is also higher than what we got before with R. The lesson that I learned here is the data training speed is much higher in Python than in R. When working with this project with R, even with only using 6000 sample data to train the model, it still takes lots of time for processing the task. However, this task can be executed with Python in a faster way with the whole data set and it helps us to get the higher performance matrix. Thus, using Python for this task is more efficient than R.