Student Name: Luyang Ye

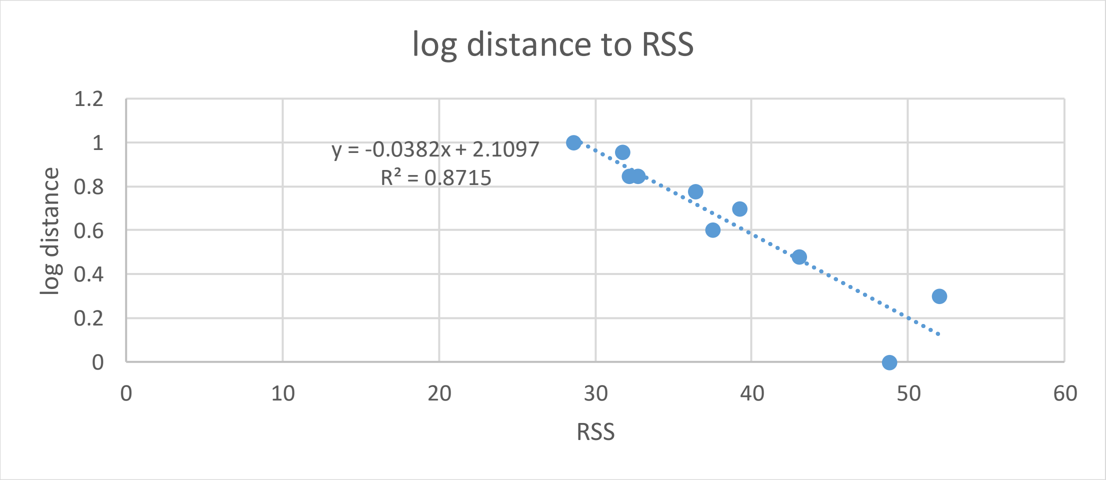
Student ID: z5280537

Project Stage 3

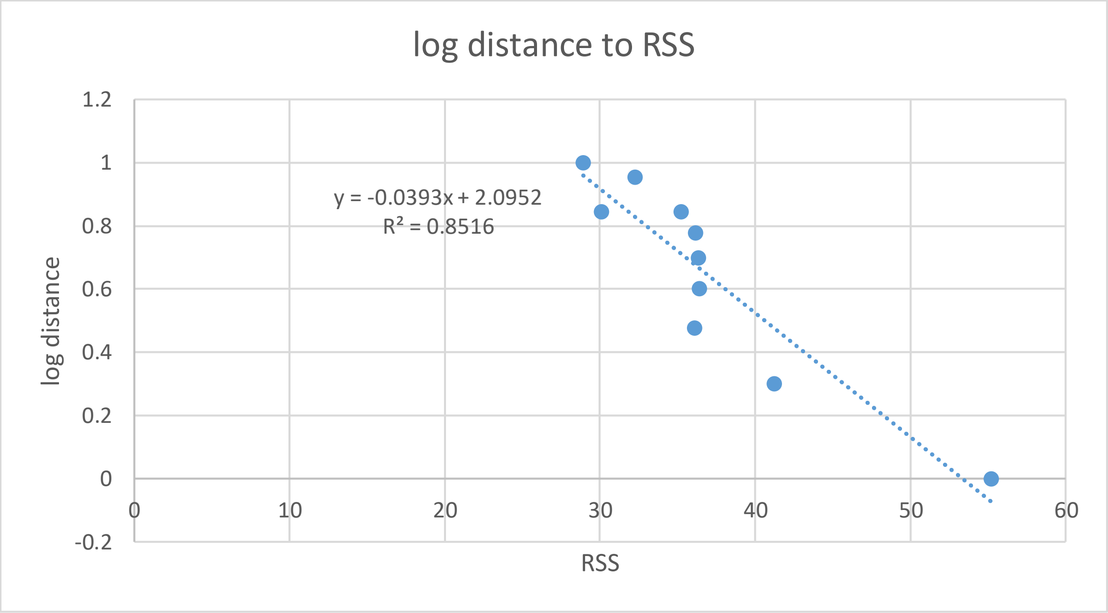
Data Set

The date is collected by an iPhone 6s plus smart phone and an ASUS laptop with windows system. The data are collected by using Microsoft Network Monitor. The outdoor data was collected at the aisle in front of the library and the indoor data was collected at the aisle in an apartment building.

The following graphs uses the average RSS as the x axis and the log distance as the y axis, the line and the formula shown on the graph is produced by linear regression.

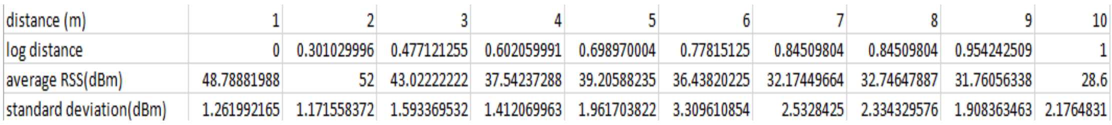


Graph 1. indoor average RSS with log distance



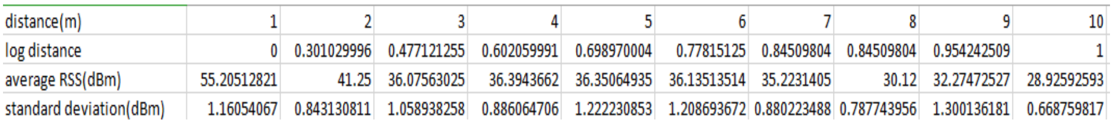
Graph 2. outdoor average RSS with log distance

The following shows some parameters that can describe the data set:



distance (m)	1	2	3	4	5	6	7	8	9	10
log distance	0	0.301029996	0.477121255	0.602059991	0.698970004	0.77815125	0.84509804	0.84509804	0.954242509	1
average RSS(dBm)	48.78881988	52	43.02222222	37.54237288	39.20588235	36.43820225	32.17449664	32.74647887	31.76056338	28.6
standard deviation(dBm)	1.261992165	1.171558372	1.593369532	1.412069963	1.961703822	3.309610854	2.5328425	2.334329576	1.908363463	2.1764831

Graph 3. parameters of indoor data set



distance(m)	1	2	3	4	5	6	7	8	9	10
log distance	0	0.301029996	0.477121255	0.602059991	0.698970004	0.77815125	0.84509804	0.84509804	0.954242509	1
average RSS(dBm)	55.20512821	41.25	36.07563025	36.3943662	36.35064935	36.13513514	35.2231405	30.12	32.27472527	28.92592593
standard deviation(dBm)	1.16054067	0.843130811	1.058938258	0.886064706	1.222230853	1.208693672	0.880223488	0.787743956	1.300136181	0.668759817

Graph 4. parameters of outdoor data set

Assume the data are in normal distribution, then the three standard deviation on either side of the average mean value will cover 99.7% possibilities of the data.

According to graph 3 and graph 4, for the indoor data set, the average standard deviation of each distance is approximately 1.9, the minimum standard deviation is about 1.2 and the maximum standard deviation is about 3.3. Considering that the average RSS for each location is at least 28, it is reasonable to regard the indoor data set as good.

Also, for the outdoor data set, the average standard deviation is approximately 1, the minimum standard deviation is about 0.67, and the maximum standard deviation is about 1.3. Considering that the average RSS for each location is at least 28, the outdoor data set is even better than the indoor data set.

Stage 1 Algorithm

The stage 1 algorithm is linear regression.

To be more specifically, this algorithm will use excel to apply a linear trend line to average RSS and the log distance. This trend line should represent the movement of RSS corresponding to the distance. Then there will be a formula for this trend line. According to the data set described above, for the indoor environment, the equation is y=-0.0382x + 2.1097, and for the outdoor environment, the equation is y=0.0393x + 2.0952(Where x is the RSS and y is the log distance).

Then these formulas can be used to estimate the log distance based on the RSS, and therefore the distance can be calculated based on the log distance.

Error Report for stage 1 algorithm

The following matrix is produced by randomly pick 182 samples, then compare the results produced by stage 1 algorithm with the real distance.

	Precision	recall	f1-score	support
1	0	0	0	13
10	0.36	0.40	0.38	10
2	0.33	0.17	0.22	12
3	0	0	0	23
4	0	0	0	41
5	0.05	0.05	0.05	20
6	0	0	0	30
7	0	0	0	12
8	0.11	0.45	0.18	11
9	0.56	0.50	0.53	10
accuracy			0.13	182
Macro avg	0.11	0.12	0.10	182
Weighted avg	0.09	0.09	0.08	182

Table 1 Matrix for indoor with linear regression

	Precision	recall	f1-score	support
1	0	0	0	15
10	0.76	0.93	0.84	14
2	0	0	0	11

3	0	0	0	17
4	0	0	0	17
5	0	0	0	12
6	0.38	0.43	0.40	14
7	0.43	0.90	0.58	29
8	0	0	0	34
9	0.06	0.11	0.08	19
accuracy			0.26	182
Macro avg	0.14	0.20	0.16	182
Weighted avg	0.16	0.26	0.20	182

Table 2 Matrix for outdoor with linear regression

According to table 1 and table 2, the overall accuracy of stage 1 algorithm for indoor environment is about 0.13, and for outdoor environment is about 0.26. The distribution of the precision for each distance is irregular. Therefore the performance of stage 1 algorithm is not very good.

After reflecting of the stage 1 algorithm, there most important reason for the poor performance of stage 1 algorithm is that although linear regression can be used to model the relationship between a dependent variable and a independent variable, it only looks at the mean of the dependent variable. In other words, since the stage 1 algorithm only used the average RSS of each distance rather than the whole data set, the mean cannot describe the whole data set. As a result, the performance of this algorithm is poor.

Stage 3 Algorithm

To improve the poor performance of stage 1 algorithm, the stage 3 algorithm have to consider the whole data set. Therefore, after some research, the MNB model is chosen. The full name of the MNB model is Multinomial Naïve Bayes. Basically, it is a kind of Naive Bayes classifier used for multinomial models. The Naive Bayes classifiers assume the value of a particular feature is independent of the value of other features, and try to predict the outcome based on the features. For this project, the different RSS values will become the features, and the distance is the outcome the algorithm try to predict. Because the possible outcomes are more than 2, the multinomial model is used.

To be more specifically, first the algorithm need a programmer to input the RSS values of the corresponding environment as features. Then the programmer need to input the distance, in other words, the outcome of each corresponding RSS values. These RSS values and corresponding outcomes is the train set, the algorithm can produce a model based on the train set. After the model is created, it can be used to predict the distance for any RSS values.

Error Report for stage 2 algorithm

To test the stage 3 algorithm, 182 samples are randomly picked from the indoor environment and the outdoor environment. The following tables are the result matrices:

	Precision	recall	f1-score	support
1	0.87	1	0.93	13
10	0.56	0.90	0.69	10
2	1	0.83	0.91	12
3	1	0.61	0.76	23
4	0.66	0.98	0.78	41
5	0.25	0.15	0.19	20
6	0.90	0.90	0.90	30
7	0.45	0.83	0.59	12
8	0	0	0	11
9	0.5	0.1	0.17	10
accuracy			0.70	182
Macro avg	0.62	0.63	0.59	182
Weighted avg	0.67	0.70	0.65	182

Table 3 Matrix for indoor with MNB

	Precision	recall	f1-score	support
1	1	1	1	15
10	0.76	0.93	0.84	14
2	1	1	1	11
3	0	0	0	17
4	0	0	0	17
5	1	0.08	0.15	12
6	0	0	0	14

7	0.40	0.72	0.52	29
8	0.91	0.88	0.90	34
9	0.87	0.68	0.76	19
accuracy			0.57	182
Macro avg	0.59	0.53	0.52	182
Weighted avg	0.59	0.57	0.55	182

Table 4 Matrix for outdoor with MNB

According to the precision row of table 3, for the indoor environment , it seems the closer the distance is, the higher the precision. For instance, while the distance is less than 4, the precision is higher than 80%, and the lowest precision is around 8 meters.

According to the table 4, for the outdoor environment, not only the overall accuracy is lower than the indoor environment, the distribution of the precision is also irregular. Which means the performance of stage 3 algorithm in the indoor environment is much better than in the outdoor environment. The possible reason of this phenomenon may be there are more factors can influence the RSS in outdoor environment. This may be solved by collect more data and built a more comprehensive model.

According to the above table, the overall accuracy of the algorithm in indoor environment is around 70%; the overall accuracy of the algorithm in outdoor environment is around 57%. Compared to the stage 1 algorithm, the accuracy of the indoor environment has improved for about 500%, and the accuracy of the outdoor environment also improved for about 200%. Obviously, the performance of the stage 3 algorithm is much better than the stage 1 algorithm.

I personally think there are two most important reasons for these improvements. The first reason is that the stage 3 algorithm has considered the whole data set rather than the mean RSS of each distance. Another reason is that, the distribution of the data is not linear. According to graph 1 and graph 2, even the distribution of the average RSS is not close to a straight line. Therefore the performance of the linear regression is not good. But the non-linear distribution of the data will not significantly influence the MNB model, because the MNB model use features to predict the outcome. As a result, these improvements are achieved.