Assessment 2 Introduction

Assessment 2 asked our group to look at the MACCDC2012 data set and then:

1. Choose an appropriate inference goal;
2. Create an appropriate baseline model
3. Use an appropriate strategy to learn any parameters of the non-trivial model
4. Use an appropriate strategy to learn about out-of-sample performance of the non-trivial model
5. Compare performance of the non-trivial model to the baseline

The report is divided into the following sections:

1. 01 – Introduction
2. 02 – K-Means Clustering
3. 03 – DBSCAN Clustering 1
4. 04 – DBSCAN Clustering 2
5. 05 – Performance Analysis
6. 06 – Conclusion

In this report, we chose to look at the missing values in the duration, orig_bytes and resp_bytes and two methods to impute them. Our baseline model is K-means clustering which allowed us to impute the values based on this algorithm. The code run in this section was created by Alex. This was chosen as it was simple to work with on this data set but also was complex enough to allow us to create some division in the data set and impute 'accurately'. In the following two sections of the report, the non-trivial model is implemented. The non-trivial model we chose was DBSCAN clustering. We assumed that DBSCAN clustering would outperform K-means clustering on this data set and therefore imputation through this method would be better. The first DBSCAN clustering is performed by Matt and the second is performed by Wenqi. In the final two sections of the report, we'll look at how well the clustering algorithms performed and use the Mahalnobis distance to create a performance metric to allow us to review the correlations between the factors we imputed. This will give us the opportunity to compare the performance of the models and try to understand the issues faced by either of them.