Machine Learning for Networking
Seminar 1 - Analyzing a dataset

1 Dataset

We will use a WiFi dataset for this seminar. It is structured as a matrix, in which the columns are features and the rows are samples. The features included are the following:

1. **Number of STAs:** the number of devices in the network.
2. **Load:** Aggregated generated load for the whole network.
3. **Size (x):** Horizontal span of the area in metres.
4. **Size (y):** Vertical span of the area in metres.
5. **Area:** Size(x) multiplied by size(y).
6. **Minimum Contention Window:** The lowest contention window used in the network, CW is in the range {3, 7, 15, 31, 63, 127, 255, 511, 1023}.
7. **Channel width:** Networks can use channels of 20, 40, 80 and 160 MHz.
8. **Packet size:** In bits, can be in the range {4000, 6000, 8000, 10000, 12000};
9. **Maximum RSSI:** The highest received signal from the AP (dBm).
10. **Average RSSI:** The average received signal from the AP (dBm).
11. **Minimum RSSI:** The lowest received signal from the AP (dBm).
12. **Average probability of failure:** Chance of packets not being delivered.
13. **Throughput:** Aggregated data received by the devices.
14. **Average delay:** Average packet transmission time (seconds).
15. **Total airtime:** Transmission time required to successfully deliver all transmissions (over one second).
16. **Proportional airtime:** Available transmission time used by all the devices in the network (over one second).

Features 1-11 will be the input of our models, and features 12 through 16 are the parts of the network that we want to predict (i.e., the output).

2 Analysis

- Some features may be more relevant than others. Find the correlation between inputs and outputs.
- Feature importance may vary depending on the scenarios. Split the dataset based on the samples that satisfy the condition $\frac{\text{Throughput}}{\text{Load}} \geq 0.975$. Then, do the correlation study once more. Is there a meaningful difference between the two cases?
- Train two linear regression models using the split datasets (one for $\frac{\text{Throughput}}{\text{Load}} \geq 0.975$ and one for $\frac{\text{Throughput}}{\text{Load}} < 0.975$) to predict the throughput of the network (use the 11 inputs and only throughput as the output).
  - Start by splitting each dataset into a training and testing dataset (70% training - 30% testing) and train the models with the training portion.
  - Use the trained models to predict the results based on the testing dataset.
  - Calculate the error of your predictions using the Root Mean Square Error.
- Now repeat the process without splitting the dataset based on the throughput. Which case has the lowest RMSE?
- Using the correlation information, select 3 features to train another model, then test it and compare its error to the model using all the features. Could we still predict the throughput accurately?
- (Optional) Using the model specification of fitlm, train the model with a quadratic regression, can the RMSE be reduced further with this model? Can we still use less features in this case and obtain positive results?