KDD Intrusion Detection

Dinis Marques Firmino

Tools & Frameworks

- Python 3.5
- Sci-Kit Learn
- Pandas
- Matplotlib
- SAS Enterprise Miner

Exploration

- Labelling data
- SAS used for initial data exploratory analysis
- Observed statistics and distributions
- Histograms to observe class frequencies
- Box plots for initial outlier detection
- Clustering

Pre Processing

- Encoding categorical variables
- Scaling inputs between 0-1.
- Outlier removal
- Rebalancing the dataset
- Feature selection using Chi2 test.

Models

- 3 neural network models in total
  - 1 for binary classification
    - Inputs: **Count, Srv_serror_rate, Protocol, Logged_in, service, Dst_host_same_src_port_rate, Dst_host_diff_srv_rate**
  - 1 for 5 class classification
    - Inputs: **Duration, Src_Bytes, Dst_Bytes, Count, Srv_Count, Dst_host_diff_srv_rate**
  - 1 for 23 class classification
    - Inputs: **Src_Bytes, Dst_Bytes, Count, Srv_Count, Serror_rate, Dst_host_srv_diff_host_rate**
- Logistic activation function
- Stochastic gradient descent optimizer
- Best performing models had 2 hidden layers with 8 units in first and 6-7 in the second layer.

Assessment

- Metrics
  - R2
  - MSE
  - Precision for each class + overall
  - Recall for each class + overall
  - Support
- Stratified K-Fold Cross Validation
  - 3 Folds
- Confusion Matrices

Results

- Binary w/ Original dataset
  - **R2: 0.922**
  - **MSE: 0.012**
- 5 Class w/ Original dataset
  - **R2: 0.9641**
  - **MSE: 0.0073**
    - **Poor u2r accuracy**
- 23 Class w/ Original merged dataset
  - **R2: 0.9772557**
  - **MSE: 0.39601419**
    - **Good at normal, neptune, smurf, back, satan, teardrop, warezclient, n_map**
    - **Bad at buffer_overflow, ftp_write, imap, ipsweep, land, loadmodule, multihop, perl, phf, rootkit, spy, portsweep, and warezmaster**