Task 2: Feature Selection Using Information Gain

Diana Pham

Progress Report:

For Task 2 of my final project, I conducted feature selection on the training dataset by using information gain as the criteria. To do so, I followed the equations that were given in the research paper and created two functions that calculate entropy and information gain for each variable. After storing the information gains associated with each variable within a dictionary, I sorted them in descending order and created feature groups according to their weights. A total of seven groups were created for features with: 1) information gain greater than 0.6, 2) information gain greater than 0.5, 3) information gain greater than 0.4, 4) information gain greater than 0.3, 5) information gain greater than 0.2, 6) information gain greater than 0.1, and 7) all features. These groups are not mutually exclusive, so if a feature was greater than 0.6, then it was also included in Group 1 as well as the other six groups.

Challenges:

While following the implementation of the original research paper titled, CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection, I ran into a few issues that led to my results differing from the research paper's results:

1. The research paper implemented feature selection using information gain in Weka software instead of python so I had to look up alternative functions that were similar to Weka's `feature_rank` function. After researching different functions, the most similar function was `mutual_info_classif` from sklearn.
2. Several of the values in the training dataset were infinity or larger than python's float64 maximum number so `mutual_info_classif` was returning a ValueError. Due to this issue, I created my own functions that calculate entropy and information gain by following the equations that were given in the paper.
3. The authors didn't specify their seed number in their paper, so the random selection of 20% of the data from the pre-processing step may be different from theirs, and as a result, the entropy and information gain values will be different.