CSCI 5622 - Machine Learning - Project Proposal

1. **Project**
    a. **Task:** We will compare the effectiveness of supervised versus unsupervised classifiers on online networking data. We will use the classifiers to predict the category type of the data (e.g. browser data, peer to peer data, monitoring tools, etc.). This could be a useful application in networking systems as this could allow you to block types of networking traffic when they originate from unusual sources (e.g. embedded medical equipment accessing bittorrent data).
    b. **Data:** We will be collecting data from local machines using Wireshark to sniff all incoming and outcoming packets. Wireshark's data includes information from all layers of the OSI model. Wireshark does not provide information on the traffic category. Therefore, in order to create labeled data, we will isolate a particular category and collect data from that time period (i.e. SSH into a machine, sniff the packets and label it as "SSH data"). Due to the robustness of Wireshark, we will be able to collect a wide range of features to use as part of our classification model.
    c. **Baselines:** Our baseline will be established as the classifier accuracy on a holdout or validation set. We will run our proposed classifier without any features or applied feature engineering to evaluate the raw predictive value of said classifier.
2. **Team**
    a. Byron Becker, Saim Khan, Kevin Holligan, Victoria Slattum, Ishita Srivastava
3. **Techniques**
    a. We will be exploring both supervised and unsupervised learning techniques for classification. This will depend on the amount of labelled data we are able to retrieve (less labelled data will encouraged an unsupervised approach, whereas sufficient labelled data will favor a supervised approach). Some of the options for possible supervised learning techniques are C4.5 decision trees, Random Forest (RF), and Bayesian Networks. For unsupervised learning, we want to explore K-means, expectation maximization, and DBSCAN.

4. **Milestones**

| Task | Estimated Date Complete |
|---|---|
| Collect and label data | March 19 |
| Preliminary Baseline Model Complete | April 2 |
| Baseline Model Complete | April 10 |
| Baseline Write Up | April 13 |
| Final Model Complete | April 23 |
| Project Write Up and Video Complete | May 4 |