Ertan Can Güner
60610

COMP430 Homework 3

Part 1

1) a) It does not change the other words doesn't change the score of another word because when we look at the score function in Section 2.3 words are used in counting functions. Score function uses words independent of each other to calculate the number of occurrences in Spam and Ham emails.

b) The only thing affects the model is to include words or not include them. We use a probability distribution vector to uniformly include these attack words to payload. There can be two approaches to this as mentioned in the paper. We can include all the words in the dictionary of the victims used language in the emails and this will drop the accuracy of the spam filter significantly especially if we include the most used words data to the dictionary, which will enable almost infiltration to both ways (false positives and false negatives) depending on the attacker motives. Other approach is to use of a more focused attack where setting the values in p vector to 1, stating that these words must be included in the attack vector, making the adversary select which narrowed category of emails should be labeled as spam.

c) RONI defense takes samples from training and test data sets, 20 and 50 messages respectively. Then it trains with the training set and queries each data on that training set to measure its impact on the classification error. If the impact of the query is too large then it is removed from the training set. Dynamic Threshold takes another approach by changing the threshold values to classify the test data. They divide the training set into two and create model from the first half where the second half becomes the test data. They obtain scores from the test data and calculate threshold values as follows: They count the number of spam emails which have lower value than t and divide that with the same number summed with number of ham emails which have higher value than t. This allows for quickly computing the threshold values with the set that we have other than using the default values, but this defense fails to achieve a defense for attacks. It never mislabeled ham messages as spam but data labeled as unsure increased heavily, it classified every spam mail as unsure.

d) We can use outlier detection to detect outliers and like RONI we can remove them from the data set.

2) a) S: set of all possible manipulations, s: subset of possible manipulations, F(.): objective function, f(.): classification output, x: data to be manipulated, $\oplus$: Manipulation function, $\check{\lambda}$: tuning parameters to

adjust tradeoff between score and penalty, C(.): penalty function, q: query threshold, T: query budget

This formulation tries to find the best (smallest) amount of perturbration to be done on the file to evade the model.

b) Query-efficient means that formulated program tries to save time and resources by not going through nonsens data to test the model, it goes through its own domain. Functionality-preserving means that their approach doesn't manipulate the file to alter its functionality rather it messes with its metadata.

c)

d) They don't leverage the concept of transferability because the papers approach managed to find that they evade more than 12 commercially available antivirus software.

Part 2

1) We can easily see that when we increase the size of the flipped data set the accuracy of the model drops significantly. Which is the expected result because we increase the false information given to the model so it will be predicting wrongly. It seems that not all 3 ML models are affected like each other. Support Vector Machines are the most resilient against these kinds of attacks. When we set n large enough SVC is the least model that dropped the accuracy so we can say that it is more robust than other models.
2) We set the threshold value to calculate the probability of the sample to determine whether the sample is from the training set. When we set the threshold value high, we observe the recall value to be small because predicted probability of the sample belonging to each class varies upon the given probability pairs (because this is a binary classification). If the value is higher than the given threshold, we can infer that the model is confident about this sample so we can say that it is most probably from the training set.

t	0.99	0.98	0.96	0.8	0.7	0.5
Recall	0.43	0.52	0.59	0.81	0.93	1.0

3) I tried to come up with a real-world example where we decide which features are can be adjusted by an adversary. I have decided that most easy metrics that can

be adjusted are Temperature and Wind Speed features. Given that I know the max and min values of the training set I can make a trigger pattern that is larger than the training data set. Where I randomly set the values for Temperature between 45-60 and for Wind Speed i randomly set them between 30-50 and for the rest, I set them between 0-10.

I generated 100 samples to be tested against the model where I generated the data to be tested a little lower than the injected samples 45-50 and 30-40 and the rest between 0-20. I expect them to be all classified from class 1 so I predicted their values from the model and compared each one to the expected label. The percentage came from there gave me the success rate of the injection.

DT num_of_samples	0	1	3	5	10
Success	0.0	0.19	0.05	0.87	0.99
LR num_of_samples	0	1	3	5	10
Success	0.0	0.65	0.87	0.99	0.9
SVC num_of_samples	0	1	3	5	10
Success	0.0	0.0	0.0	0.0	1.0

4) I have decided to implement a greedy algorithm to go through all the possible combinations of features to add noise to and generate a distribution based on the combination to add noise to the data. I increase the noise each time until the label flips. It currently generates 5000 distributions for each combination and increases the amount 0.05 each time.

   Avg. perturbation for DT : 1.5212

   Avg. perturbation for LR : 10.0151

   Avg. perturbation for SVC : 6.7299

5) I found data evaded left hand side and tested it on the top models. As it can be seen data that evaded SVC had the highest transferability to other models.

	DT	LR	SVC
DT	X	%30	%22.5
LR	%22.5	X	%17.5
SVC	%40	%62.5	X