Ali OKTAY

72007

Homework 3 Report

PART 1)

Question 1)

Accuracy of Model	DT	LR	SVC
0.05	0.967	0.979	0.953
0.1	0.957	0.975	0.950
0.2	0.933	0.971	0.946
0.4	0.784	0.956	0.759

Because it changes more the properties of the data as p increases, the accuracy decreases in all models as the flipping probability increases. However, not all models are affected by that modification in the same way. LR had the least amount of an impact on all of the models in my experiment, as the examples show.

Question 2)

Accuracy of Model	Correctly Identified	Total
0.05	20	48
0.1	46	96
0.2	90	192
0.4	181	384

I utilized the IsolationForest for the defense, setting the contamination at 0.5 and the number of estimators to 50. The identification increases with increasing outliers, however in such scenarios, incorrect outlier classifications become problematic. Elements that deviate from their estimators are designated as outliers by the IsolationForest. Then I verified that my outlier indexes and the flipped indexes match. My defense is not entirely successful because it only captures about 50% of the flips

Question 3)

```
Evasion attack executions:
Avg perturbation for evasion attack using DT : 0.775
Avg perturbation for evasion attack using LR : 0.896875
Avg perturbation for evasion attack using SVC : 0.8625
##################################################
```

The steps of the attacking plan are as follows: Using the trained model, it first determines the example's original class. It then copies the original sample and makes incremental changes to its features. Every feature is modified by adding and removing a predetermined increment value. It determines whether the estimated class and the original class are different after every update. An adversarial example is considered successful if the modified example yields a different predicted class after the modification. The loop keeps going until an appropriate perturbation is found or the maximum increment value is reached.

Question 4)

```
Transferability of evasion attacks:
Out of 40 adversarial examples crafted to evade DT :
-> 19 of them transfer to LR.
-> 18 of them transfer to SVC.
Out of 40 adversarial examples crafted to evade LR :
-> 18 of them transfer to DT.
-> 27 of them transfer to SVC.
Out of 40 adversarial examples crafted to evade SVC :
-> 14 of them transfer to DT.
-> 26 of them transfer to LR.
```

My results indicate that the evasion attacks have a high cross-model transferability since they may be applied to most other models.

PART 2)

By identifying non-English terms in the backdooring data, my defense aims to clean it up. In order to prevent potential backdoor attacks, it filters out non-English words from the text.

In order to accomplish this, I first built a line that determines if a word is in English or not. After that, I tokenize the string, go over each word, and pay attention to the punctuation. Then, after removing any non-English terms, join all of the strings.

I use the being of labels one to determine whether or not a string in the data has a backdoor injection and look over these rows. However, I don't reset their labels to 0 because it's possible that I removed an injection rather than a true word. We cannot guarantee that every removal removes all backdoored non-English words.

Using the table, we may draw the conclusion that when the poison rate rises, the attack becomes more successful and the defense becomes less successful since there are more poisoned data points overall. Additionally, as the trigger phrase gets longer and the number of words that are injected gets bigger, the attack becomes more difficult to intercept, increasing the attack's success rate.