

Assignment #: 3

Venkata Sai Chelagamsetty

November 12th, 2022

1. (a) The security of any system is measured with respect to the adversarial goals and capabilities that it is designed to defend against—the system’s threat model. Threat model is a model in which potential threats, such as structural vulnerabilities or the absence of appropriate safeguards, can be identified and enumerated, and countermeasures prioritized. Threat model answers questions like What kind of threats might a system face?, What kind of capabilities might we expect adversaries to have?, What are the limits on what the adversary might be able to do to us?.
- (b) An attack occurs when someone attempts to exploit a vulnerability. A compromise occurs when an attack is successful. In other words, when a compromise occurs, a vulnerability is exploited.
- (c) A security policy is a document that expresses clearly and concisely what the protection mechanisms are to achieve. It is a statement of the security we expect the system to enforce. A security model is a specification of a security policy:
it describes the entities governed by the policy,
it states the rules that constitute the policy.
Essentially It ensures that the CIA-Triad is maintained. It answers questions like how data is accessed? What level of security is required? and what should happen when these requirements aren’t met.
- (d) Confidentiality ensures that computer-related assets are accessed only by authorized parties
Integrity requires that computer system assets and transmitted information be capable of modification only by authorized parties
Availability: The degree to which data or systems are accessible and in functioning condition.
- (e) In ML systems, in terms of confidentiality, the attacks are classified as
Model Extraction : An adversary aims to discover the structure or parameters of the model by observing its predictions.
Whereas in terms of Integrity, the attacks are classified as
Poisoning attacks : An adversary tries to manipulate the training dataset in order to control the prediction behavior of a trained model such that the model will label malicious examples into a desired classes e.g., labeling spam e-mails as safe).
and , evasion attacks: The attacker manipulates input samples at test time to evade (cause a misclassification) a trained classifier at test time.
- (f) i. $Accuracy = 100 * \frac{Attacksclassifiedasattacks+Benignclassifiedasbenign}{Totalno.oflogins}$
 $Accuracy = 100 * \frac{121806}{130200}$
 $Accuracy = 93$
ii. $P(attack) = \frac{No.ofattacks}{Totallogins}$
 $P(Attack) = \frac{11250}{130200}$
 $P(attack) = 0.0864$

iii.

$$P(flag|attack) = \frac{Attackclassifiedasattack}{Totalactualattacks} \quad (1)$$

$$P(flag|attack) = \frac{11086}{11250} \quad (2)$$

$$P(flag|attack) = 0.9854 \quad (3)$$

iv.

$$P(flag) = \frac{trafficpredictedasattack}{Totallogins} \quad (4)$$

$$P(flag) = \frac{20036}{130200} \quad (5)$$

$$P(flag) = 0.1538 \quad (6)$$

v.

$$P(benign|flag) = \frac{P(flag|benign)P(benign)}{P(flag)} \quad (7)$$

$$P(benign) = 1 - P(attack) \quad (8)$$

$$P(benign) = 0.9136 \quad (9)$$

$$P(flag|benign) = \frac{Benignclassifiedasattack}{Totalactualbenign} \quad (10)$$

$$P(flag|benign) = \frac{8950}{110000 + 8950} \quad (11)$$

$$P(flag|benign) = 0.07524 \quad (12)$$

$$P(benign|flag) = \frac{0.07524 * 0.9136}{0.1538} \quad (13)$$

$$P(benign|flag) = 0.4469 \quad (14)$$

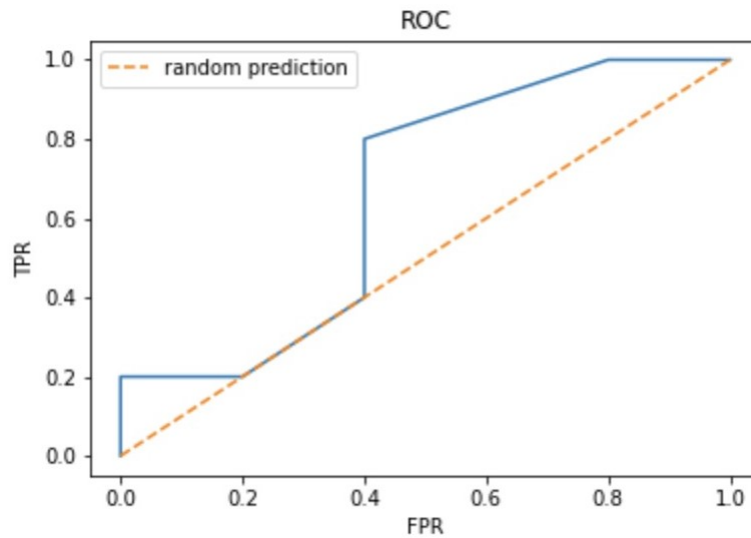
2. (a) The simplest and most accurate score function would be : $2a_1 + a_2 - a_3$, If the score goes above 10, then the score is 10. If the score goes below 0, then the score is 0. (As attribute a_1 is correlated twice as strongly with SPAM as a_2 , and Attribute a_3 is negatively correlated with SPAM exactly as strongly as a_3 is positively correlated with SPAM.

- (b) The scores would be as follows:

M1	Spam	4
M2	Not Spam	1
M3	Spam	9
M4	Spam	2
M5	Not Spam	7
M6	Not Spam	2
M7	Not Spam	8
M8	Spam	3
M9	Spam	7
M10	Not Spam	2

	Spam	Not Spam
(c) Spam	4	2
Not Spam	1	3

- i. False positive rate: $\frac{2}{5} = 40\%$
- ii. True positive rate: $\frac{4}{5} = 80\%$
- iii. False negative rate: $\frac{1}{5} = 20\%$
- iv. True negative rate: $\frac{3}{5} = 60\%$



- (d) ROC is above.

3. (a) The size of the convoluted matrix is 22×22
- (b) We get the following 3×3 matrix :

4	6	5
6	6	8
9	8	8

4. Code
5. I used a subset of first 100 samples from the test dataset and used all the four methods given in the code i.e FGSM, Basic iterative method, Saliency map, and Universal Perturbation. The model summary is given in the code.
The accuracies seem to be really good without much tuning, though I have tested 2 different CNN models and implemented the better one in that case.
Finally, we can infer that the CNN and ANN models are trained on a dataset, but when we generate the adversarial samples, we basically generate data that has different distribution compared to the test and training set. Hence the CNN and ANN models perform so poorly. When we have the augmented dataset, almost half the dataset comes from the same distribution as the training dataset, therefore the accuracies are nearly 50 in such cases. And as the test data is from the same distribution, accuracy is very high.