CS 6431: Security and Privacy Technologies
Homework #6

Due: May 18, 2018

Instructors: Tom Ristenpart and Vitaly Shmatikov

Problem 1: The Ray-Ligatti definition of code injection is: "A taint-tracking application outputs a program that contains a symbol that is both injected and code."

The NIE property is: "Removing all injected symbols from an output program should only delete or contract non-code tokens."

(a) Prove that if a code injection occurs, the program must violate the NIE property.

(b) Write a program that violates the NIE property but does not satisfy the Ray-Ligatti definition of code injection.

(c) Does the program you wrote in **(b)** have a security vulnerability? Explain why or why not.

Problem 2: In an evasion attack there is some target machine learning prediction function f which, for simplicity, we assume takes an input x and outputs a class label in $\{0, 1\}$. An evasion algorithm $\mathcal{A}$ takes as input a value x^* and seeks to generate a value x' that is "close" to x^*, yet for which $f(x^*) \neq f(x')$. Here close can be defined in many ways, in the Srndic and Laskov paper it was defined as being parsed equivalently.

(a) The Snrdic-Laskov work assumes adversarial access to similarly-distributed, labeled training data. Assume instead that the adversary only has a uniform prior (meaning, it knows nothing about the distribution of data from which the target model was trained). In theory, one could use model extraction as explored by Lowd and Meek and Tramer et al. (`https://arxiv.org/`

`pdf/1609.02943v2.pdf`) to perform attacks. Describe an approach based on model extraction and describe in detail the experiments that would need to be performed to evaluate the feasibility of such an approach.

(b) The Snrdic-Laskov attack was what I'll call non-interactive: the adversary $\mathcal{A}$ was not allowed to query the prediction function f while generating their output x'. Consider instead performing adaptive attacks that can query f (in this case, PDFrate) in the so-called FC setting. Describe an adversary for the PDFrate classifier that takes advantage of adaptivity and explain why it outperforms the non-adaptive adversary of Snrdic and Laskov.

Problem 3: Given access to an ML model f and a value (x, y), membership inference tasks an adversary with predicting whether (x, y) was in the training set used to learn f. Given access to an ML model f and a value (x, y) with some features of x removed, model inversion tasks an adversary with learning all of x.

(a) Describe a situation in which model inversion would work equally well for values (x, y) both taken from the training set and that are independent from it. Describe a situation in which model inversion would work well only for values (x, y) taken from the training set. State any assumptions you make about the model f, the distribution of (x, y) values, and adversarial access (white-box or black-box) for both scenarios.

(b) Build a membership inference attack that uses model inversion to infer membership. Referring to your two scenarios, discuss when a model-inversion-based attack would work well for membership inference.

(c) Discuss whether differential privacy mechanisms are suitable for use in protecting against model inversion attacks. What about membership inference attacks?

(d) (Bonus) Derive an upper bound on the differential privacy budget ϵ that would ensure membership inference accuracy (as measured in the Shokri et al. paper) is at most 0.55. Find a recent paper on differential privacy for machine learning models that give empirical results for different values of ϵ. Do their chosen values of ϵ satisfy your upper bound?