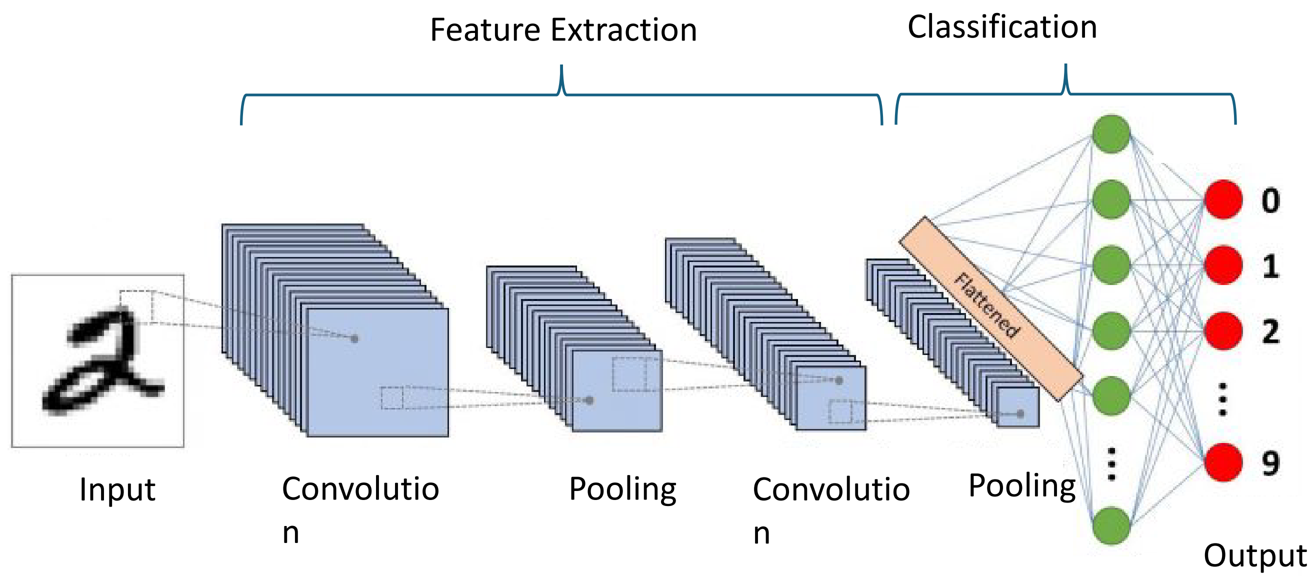
An AI Perspective for Attacks

Convolutional Neural Network

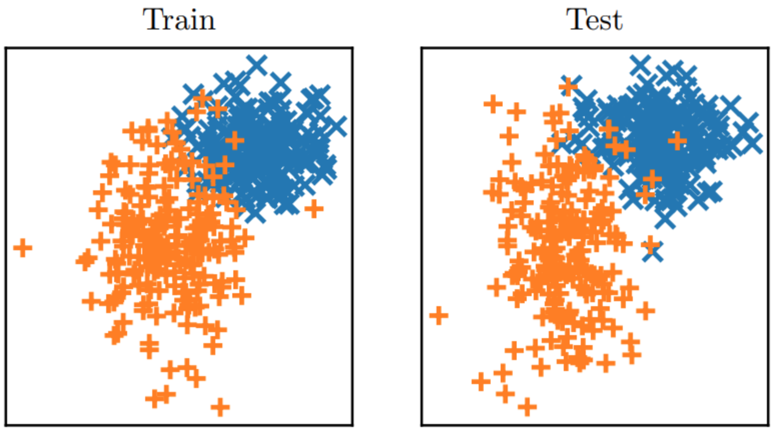


The fundamental architecture that most of the state-of-art machine learning techniques based on

The assumption for most ML techniques

All train and test examples drawn independently from the same distribution, i.e.,

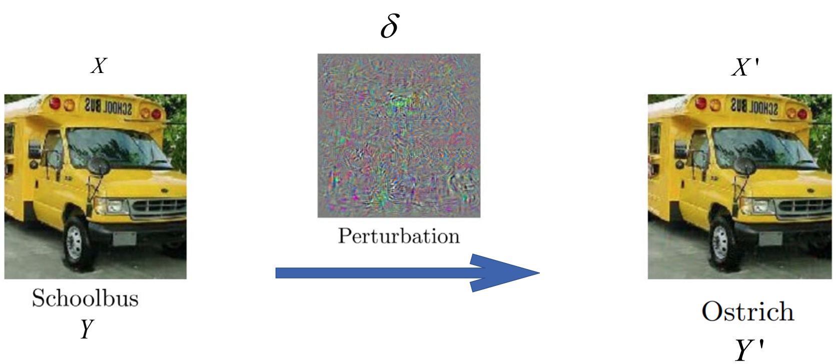
I: Independent
I: Identically
D: Distributed



Adversaries may supply data that violates that statistical assumption!

Adversarial samples

Adversarial examples are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake (Goodfellow et al 2017).



(Szegedy et al. 2013)

$$X \rightarrow Y,$$

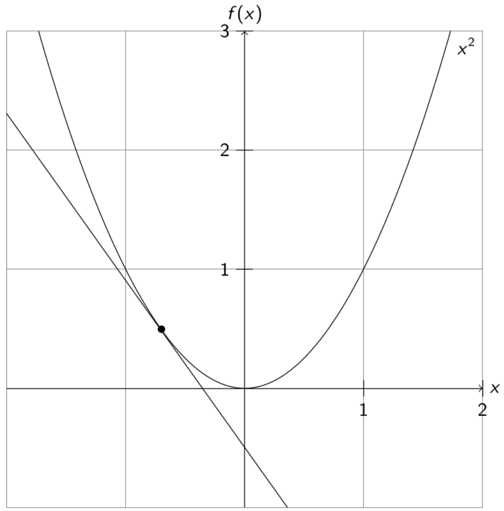
$$X + \delta = X',$$

$$X' \rightarrow Y',$$

$$Y' \neq Y.$$

Basic idea of *fast Gradient Sign Method*

- **Gradient descent** is an optimization method that can be used to find the local minimum of a differentiable function.
- If we want to find the local minimum of f(x), we first pick an initial value of x, and then compute the derivative of f(x) according to x and evaluate the initial guess. Based on the sign of the slope, which is the derivative of f(x), we can know whether we should increase or decrease x to decrease f(x).



Basic idea of *fast gradient Sign Method*

- We can use the gradient descent to train the machine learning model.
- Suppose we want to train a lineal model y = ax+b, in which a, and b is the parameter we want to train.
- Based on the idea of gradient descent, we define the loss function as:

$$L(x, y, a, b) = (y - (ax + b))^2$$

- The loss is the squared difference of real value y and ax+b, which is the prediction.
- In order to train the model in terms of a and b, and minimize the loss function, we update a and b through the slopes of the loss function regarding to a and b respectively.

$$\frac{\mathrm{d}L}{\mathrm{d}a} = 2x(ax + b - y) \qquad \frac{\mathrm{d}L}{\mathrm{d}b} = 2(ax + b - y)$$

Basic idea of *fast gradient Sign Method*

- What does it mean if we compute the derivative of the loss function with regarding to x?

$$\frac{\mathrm{d}L}{\mathrm{d}x} = 2a(ax + b - y)$$

- It can be used to make our changes of x in a way that is not obviously detected by an observer. As we found the fastest direction to change x.

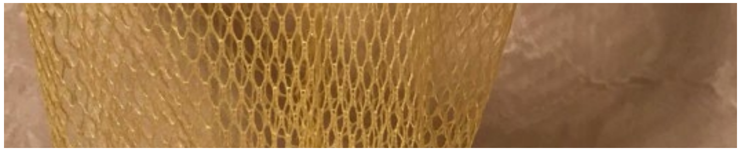
- Therefore, the adversarial perturbation is denoted as:

$$\eta = \varepsilon \ \mathrm{sign}(\nabla_x L(\theta, x, y))$$

and the final adversarial sample is denoted as $x_{adv} = x + \eta$

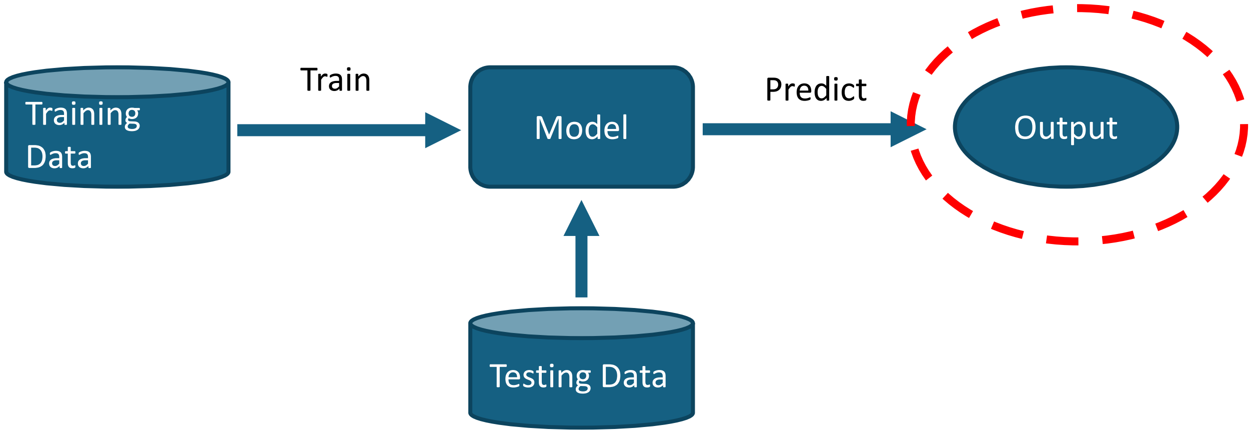
Adversarial samples: norm ball

- Adversary perturbs points within



When a vulnerability is found, the attacker can repeatedly send a single mistake to launch the attack, a.k.a. test set attack.

Adversarial machine learning

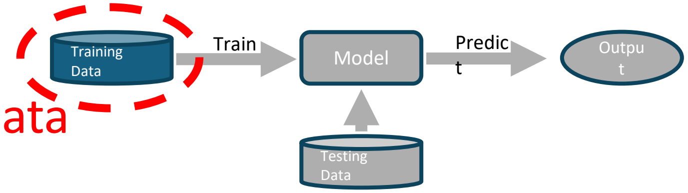


Adversarial machine learning is a machine learning domain that involves fooling models by supplying deceptive inputs.

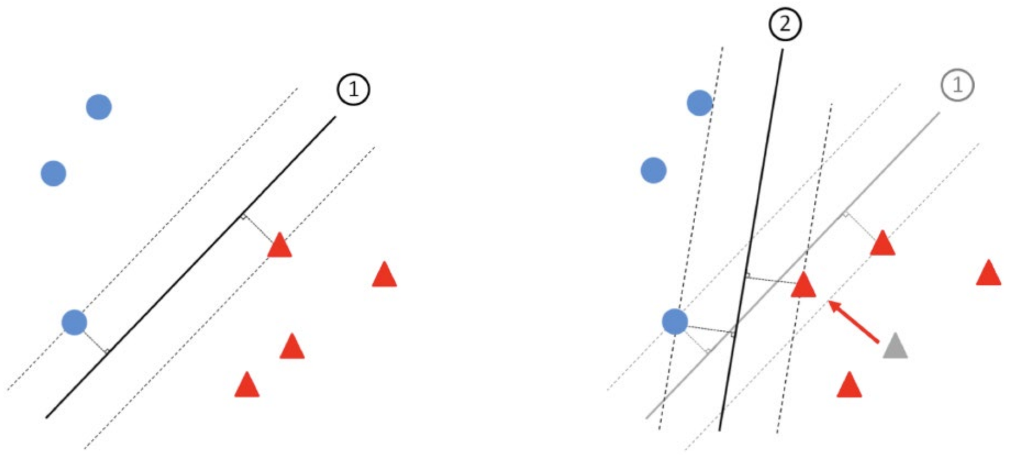
A broader attack perspective against ML

- Black box attack
- White box attack

A broader attack perspective against ML



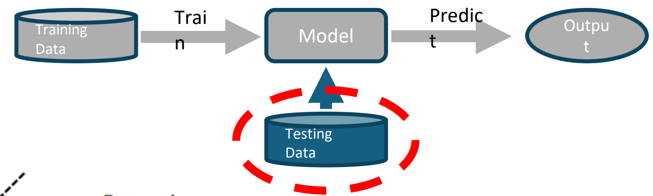
- Poisoning
  - Adversarial contamination of training data

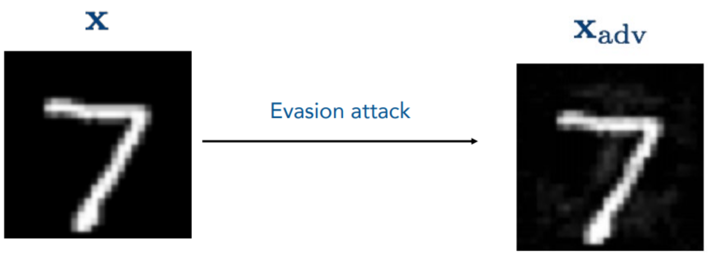


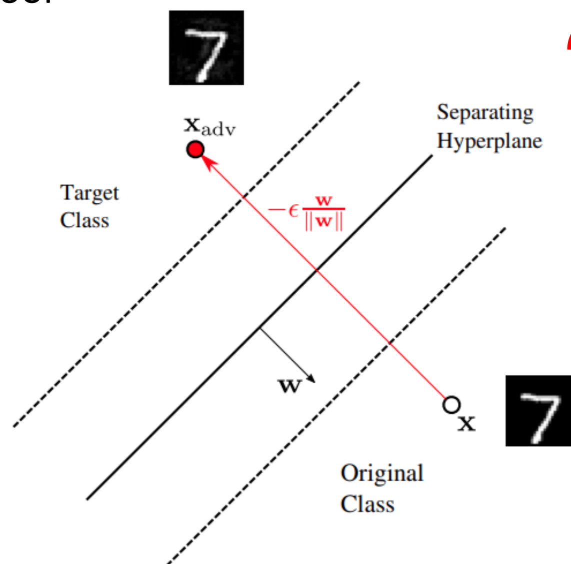
The decision boundary of SVM can be changed by just modifying one data point.

A broader attack perspective against ML

- Evasion
  - Carefully perturbing testing inputs to fool the learned model.

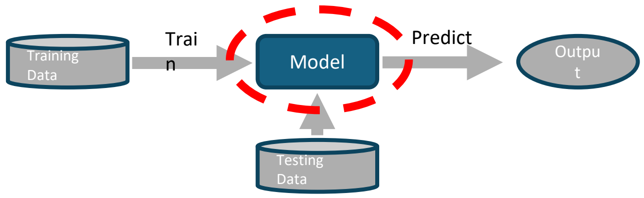




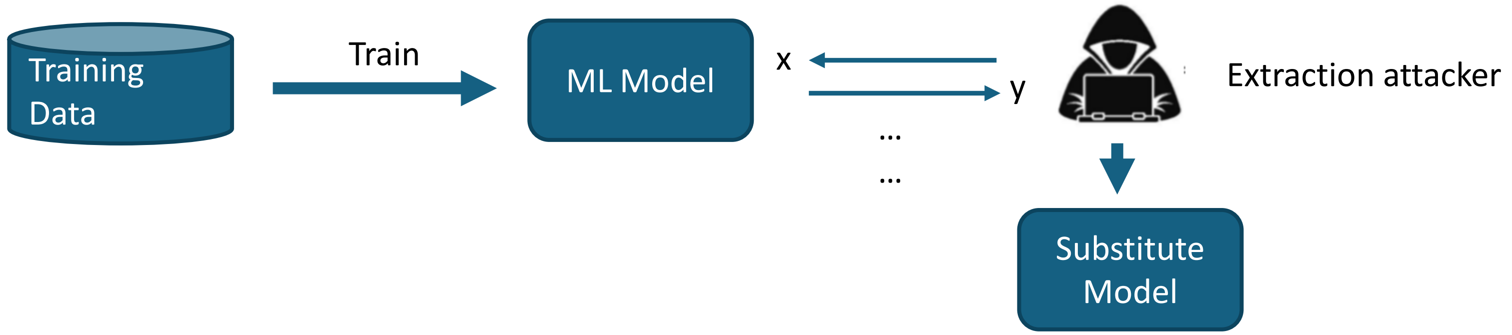


Attackers can find a perturbation of the input that moves the input cross the decision boundary.

A broader attack perspective against ML



- Model extraction
  - Gradually train a substitute model that reproduces the predictive behavior of the target model through black-box access.

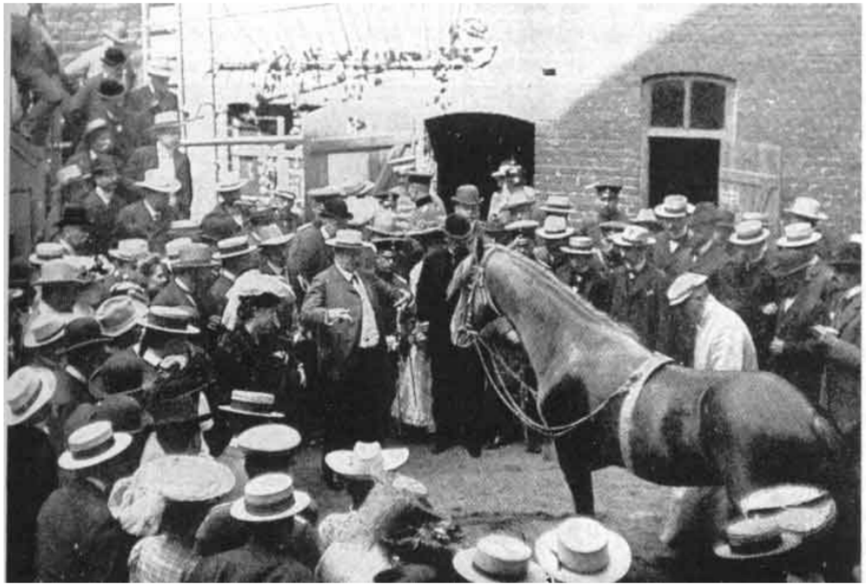


The learned substitute model can be used to generate adversarial samples or predict outputs of the targeted model.

Other challenges of machine learning

- Privacy
- Transparency
- Fairness
- Accountability
- Unlearning
- ……

Clever Hans



When cybersecurity meets adversarial ML

- Machine learning has become a vital technology for cybersecurity.
- Machine learning preemptively stamps out cyber threats and bolsters security infrastructure through pattern detection, real-time cyber crime mapping and so on.
- When machine learning techniques are widely deployed, challenges of machine learning are challenges for all!